DOCUMENT RESUME

ED 135 204 FL 008 207

AUTHOR Leslie, Adrian R.
TITLE Quest for a Computerised Semantics.
PUB DATE Apr 73
NOTE 181p.; Master's Thesis, University of Victoria

EDRS PRICE MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS Applied Linguistics; *Computational Linguistics; Computers; Computer Science; Deep Structure; *Dictionaries; Grammar; *Information Retrieval; Lexicography; Linguistic Theory; *Machine Translation; *Semantics; Statistics; Surface Structure; Syntax; Transformation Generative Grammar; Transformation Theory (Language)

ABSTRACT

The objective of this thesis was to colligate the various strands of research in the literature of computational linguistics that have to do with the computational treatment of semantic content so as to encode it into a computerized dictionary. In chapter 1 the course of mechanical translation (1947-1960) and quantitative linguistics is traced to demonstrate the limitations of computational linguistics without semantics. Chapter 2 covers linguistic research in the 1960's, which was essentially an offshoot of transformational grammar. In chapter 3, various classification schemes are examined as a body of experience from which to draw conclusions on the constraints to which the construction of a computerized dictionary is subject. Chapter 4 is a synthesis of all this data in the form of a model dictionary entry. In chapters 2 and 3 the approaches to semantics are of two types. In one, the semantic categories for each dictionary entry were in the form of unordered elements, and the means of applying them in text was placed within the realm of grammar. In the other type, syntagmatic relationships occurred between the encoded components of dictionary definitions just as they did between those of utterances in a text. The conclusion reached is that the latter type of approach provides firmer foundations upon which to set up a computerized dictionary, as it shows how information is structured in terms of its application in text. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED135204

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

QUEST FOR A COMPUTERISED SEMANTICS

by

ADRIAN ROY LESLIE

B.A., University of Wisconsin, 1969

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in the Department

of

Linguistics

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Adrian R. Leslie
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

We accept this thesis as conforming

to the required standard

.....

.....

.....

.....

ADRIAN ROY LESLIE, 1973

UNIVERSITY OF VICTORIA

April 1973

FL008207

ABSTRACT

The objective was to colligate the various strands of research in the literature of computational linguistics that have to do with the computational treatment of semantic content so as to encode it into a computerised dictionary. In chapter 1 the course of mechanical translation (1947-1960) and quantitative linguistics is traced to demonstrate the limitations of computational linguistics without semantics. Chapter 2 covers linguistic research in the 1960's, which was essentially an offshoot of transformational grammar. In chapter 3, various classification schemes are examined as a body of experience from which to draw conclusions on the constraints to which the construction of a computerised dictionary is subject. Chapter 4 is a synthesis of all this data in the form of a model dictionary entry.

In chapters 2 and 3 the approaches to semantics are of two types. In one, the semantic categories for each dictionary entry were in the form of unordered elements, and the means of applying them in text was placed within the realm of grammar. In the other type, syntagmatic relationships occurred between the encoded components of dictionary definitions just as they did between those of utterances in a text.

The conclusion reached is that the latter type of approach provides firmer foundations upon which to set up a computerised

dictionary, as it shows how information is structured in terms of its application in text. In the model structured entry the components of the definitions of a word are tagged according to their part of speech. The representation of a discourse then is determined by mapping out the dictionary definitions of words as they occur in text.

The necessity for integrating semantic components into the structure of a dictionary was brought out at the sixth annual symposium on mechanical translation held by the National Research Council at Ottawa in April, 1972. Several teams apparently achieved some results through the parsing of texts by means of ad hoc dictionaries. Since the problems of cross-referencing and of the recognition of semantic grouping remain unsolved, the key to success even in mechanical translation lies in computerised lexicography. It seems unlikely that much more progress will be achieved without it. While the solution of these complex problems is beyond the scope of this single thesis, a formulation of them is provided as a starting point for further research.

CONTENTS

INTRODUCTION

Chapter 1 COMPUTATIONAL LINGUISTICS WITHOUT SEMANTICS, 1947-1960

1.1 The Role of Statistics in Computational Linguistics 1

1.1.1 Quantitative Linguistics 1

1.1.1.1 Literary Statistics 1

1.1.1.2 Optimal Systems of Language Structure 4

1.1.1.3 Economics of Mechanical Translation 5

1.1.2 Elicitation of Paradigmatic Relations through Statistics . 8

1.2 Information Retrieval 10

1.2.1 Derivative Indexing 12

1.2.1.1 Statistical and Positional Criteria 12

1.2.1.2 Syntactic Criteria 14

1.2.2 Assignment Indexing 15

1.2.2.1 Notional Families 16

1.2.2.2 Interlingual Approach 17

1.3 Mechanical Translation without Interlingua 20

1.3.1 Non-linguistic Models of the Translation Process 20

1.3.2 Translation by Applied Linguistic Methods 22

1.3.2.1 Computerised Lexicography 23

1.3.2.1.1 Resolution of Ambiguity 24

1.3.2.1.2 Stems and Endings 28

1.3.2.1.3 Idioglossary 30

1.3.2.2 M.I.T. School and the Application of Syntax 31

Chapter 2 STRANDS OF SEMANTIC THEORY, 1960-1972

2.1 The Notation and Functioning of the Lexicon 38

2.1.1 Source Material for a Semantic Theory 38

2.1.1.1 Distinctive Feature Analyses 39

2.1.1.2 Antonyms 41

2.1.2 Katz and Fodor's Marker Theory 42

2.1.3 Determination of the Semantic Content of the markers 45

2.1.4 The Integration of Marker Logic with Syntactic Relations 48

2.1.4.1 Ordering of the Markers 49

2.1.4.2 Modifications in the Notation for Marker Representation 53

2.2 The Semantic Interpretation of Syntactic Structure 57

2.2.1 Cases 57

2.2.2 The Structuring of Concepts via Syntax 61

2.2.3 Empirical Approaches to the Linguistic Organisation of knowledge 64

2.2.3.1 Quillian's Memory 65

2.2.3.2 Schank's Conceptual Structures 68

Chapter 3 CONTRIBUTIONS OF THE CLASSIFICATION SCHEMES

3.1 General Classifications 75

3.1.1 Drawbacks of the Enumerative Schemes 75

3.1.2 Differentiation of Analytic and Synthetic Relations 81

3.2 Special Classifications 86

3.2.1 System without a Syntax, Coordinate Retrieval 86

3.2.1.1 Methods of Organising Data in Coordinate Retrieval 86

3.2.1.2 Reliance upon Practical Criteria 89

3.2.1.3 Necessity of a Notational Syntax 93

3.2.2 Classifications having a Notational Syntax 97

3.2.2.1 Syntagmatic Relations Expressed by Relationship Items .. 98

3.2.2.2 Syntagmatic Relations Expressed by Role Indicators 101

Chapter 4 GROUNDWORK FOR A COMPUTERISED DICTIONARY

4.1 Prolegomena to a Computerised Semantics 108

4.1.1 Noel's Syntactic Tree of Cases 108

4.1.2 Wilks' Context Specifying Descriptors 116

4.2 Structure of a Comprehensive Computerised Dictionary 119

4.2.1 Overhauling of Katz and Fodor's Marker Tree 119

4.2.2 Context Specifying Descriptors in a Computerised Dictionary 123

4.2.3 Scope of a Computerised Dictionary 129

BIBLIOGRAPHY 137

LIST OF FIGURES

Figure 1: Marker Tree 43

Figure 2: Quillian's Memory 66

Figure 3: Ledley's Tabledex 88

Figure 4: Format of Mooers' Entry Card 90

Figure 5: Farradane's Operators 97

Figure 6: Noel's Syntagmatic Tree 110

Figure 7: Rangathan's Notation Expressed in a Transformational Tree 112

Figure 8 a and b: Noel's Syntagmatic Trees 114-15

Figure 9: Binary Branching of Marker Trees 122

Figure 10: Reconstructed Tree for Harrap's 'sucker' Entry 124

Figure 11: Model Tree for 'sucker' 126

Figure 12: Model Tree for 'baste' 128

Figure 13: Deep and Surface Structure Tree for 'move' 133

Figure 14: Computationally Constructed Tree 134

ACKNOWLEDGEMENTS

My greatest indebtedness is to Dr. J-P. Vinay who provided the impetus which led me to undertake this work. During the two years in which it was being developed, the benefits of his experience in computerised lexicography and his comments and suggestions have been cardinal in the structuring of the content of this thesis.

I also wish to thank Dr. A. Baartz for valuable suggestions, and Mr. B. Kallio of IBM and the various students in linguistics 481 in 1972 for healthy criticism that diverted me from an unpromising approach to computerised lexicography in the last chapter.

I am indebted to Dr. H. Huxley of the Classics Department for advice which led me from Classics to Linguistics, and to Dr. F. G. Cassidy of the University of Wisconsin for critical insights into example sentences.

I wish to thank both my parents for continuous support and criticism and my mother for typing the first draft of this thesis.

I also wish to thank Mrs. E. Richardson for typing the final draft of this thesis.

INTRODUCTION

This thesis is in support of the single approach to mechanical translation, fact retrieval and document retrieval, in which semantic content is organised into a dictionary for computation, as opposed to the separate and thus far ad hoc treatments of each. The success of non-semantic approaches in other areas of computational linguistics has been demonstrated. In stylostatistics, for example, while it is usual for a human researcher to focus on the meaning of texts of disputed authorship, it is not mandatory. The idiosyncracies of an author's style may be observed in the appearance a given number of times of certain words or morphemes or other marked indices that a machine may readily identify.

In mechanical translation and fact retrieval, too, marked indices were sought. Fact retrieval in its most rudimentary form consisted of counting words in a text and making a summary of it by extracting the most frequent words. This technique was improved by classifying words according to subject area into what were called "notational families" on the hypothesis that the subject area of a text would be revealed by an accumulation of words belonging to one area. In mechanical translation such families were named idioglossaries and were applied to disambiguate words. Such categorisations of vocabulary were an acknowledgement of the value of providing an organisation of semantic content as the general classifications of the type invented by Dewey in the last century for document retrieval do.

Another line of development consisted of representing semantic content by means of unordered descriptors. In document retrieval the special classifications were so constructed in order to make provision for more than one organisation of information so as to meet a user's needs. In the period between 1947 and 1960 unordered descriptors in the form of concept numbers were set up in mechanical translation to complement the idioglossary in disambiguation. These numbers, which were based upon statistics that indicated that the scanning of one or two words on either side of the ambiguous one would be effective, indicated the selectional restrictions that allowed words, or more specifically the meanings of words, like "flowering" and "plant" to be immediate constituents. Disambiguation would take place through the matching of such concept numbers. In the 1960's Katz and Fodor arrived at a system of markers which were essentially concept numbers factored into semantic categories.

The main drawback of applying unordered descriptors was that from a finite vocabulary of them only a finite number of combinations could be produced, whereas in natural language syntagmatic structure allowed the creation of potentially infinitely long and many sentences. Consequently the representation of texts by descriptors was precluded. In coordinate retrieval they seem to be viable only because of the limits of a library's holdings and therefore, of the number of discourses to be represented. This same principle might appear to apply to the above-mentioned markers on the ground that as the body of information contained in a dictionary or encyclopedia is finite, so is the number of semantic categories

necessary. However, the presence of syntagmatic structure in natural language weighs against it. For example, while categories of the type abstract and animate might be appropriate in some way in the dictionary entry for "frighten", they would not be useful by themselves for detecting, for example, "sincerity frightens John" as grammatical and "John frightens sincerity" as anomalous.

The key to a computerised semantics lies in explicit paraphrase in the form of syntagmatically construed elements. This type of paraphrase has been the basis for research at the Cambridge Laboratory Research Unit in the 1950's and in the 1960's and 1970's at Stanford too. Where it is carefully formulated, the necessity for explicitly stating the paradigmatic relationships between dictionary entries, such as that of hyponymy, no longer exists. In the latest research at Stanford the functions of parts of speech and descriptors have been integrated into what are called semantic elements. These are the basis for the type of dictionary structure suggested in this thesis, in which Katz and Fodor's marker tree would be reformulated.

The ramifications of establishing such a structure have not been investigated in this thesis. The structures provided for the two examples in chapter 4 serve only as illustrations and are the end product of a survey of the literature in computational linguistics, being based upon the elimination of various fruitless approaches. Such an investigation would require the alignment of the representations for computation of several words and consequently an analysis of a large amount of data. The formulation of the semantic content

of dictionary entries with sufficient rigour for testing in scale, therefore, comes within the province of a work more comprehensive than this thesis.

1 COMPUTATIONAL LINGUISTICS WITHOUT SEMANTICS, 1947-1960

Before about 1960 no attempt was made in computational linguistics to provide a model of how semantic content may be structured for computation. To solve semantic problems recourse was had to discrete indices of surface structure. In mechanical translation it was some specific word or words in the context that was sought to resolve multimeaning. In fact retrieval and in research on disputed authorship the statistical analysis of word counts replaced the complex analysis of content by humans.

1.1 The Role of Statistics in Computational Linguistics

1.1.1 It is upon the interpretation of these counts that the successful use of statistics in computational linguistics depends, for they may signify facts, for example, either about a given language or a particular writer's style. Thus a letter may recur because it is a literary device as in the case of alliteration, or because it is an affix common to a group of frequently occurring words. The study of these facts belongs to quantitative linguistics, which, to adapt Herdan's[1] divisions, may be divided into the following three branches: literary statistics (or stylostatistics), optimal systems of language structure and mechanical translation economics.

1.1.1.1 Since writers can choose their own words, the applicability of statistics to the first branch, in which authorship questions are involved, may appear surprising. But it is only the first few words

(samples) that are insufficient for statistical procedure. As more and more words come under study, the author's choice of each additional word is subject to the rules of grammar, which put constraints on the possible variations that may occur in the ratios between the occurrences of various words. The analogy which Herdan[2] draws between De Saussure's[3] "langue-parole" dichotomy and the dichotomy between population and sample is appropriate. "Parole" (the individual act of speech), like the sample, is open to individual choice, but as the number of acts of speech increases they are constrained by "langue", a body of fixed conventions. Within the limited range of variation permitted by the language the writer's own preferences for certain words form a statistical pattern. This pattern is an attribute of a given style that distinguishes it from other styles.

This quantitative approach has helped to solve such problems of literary research as the determination of the chronological order of texts although the approach has not been able to show the development of an individual author's style, and the identification of authors of hitherto anonymous texts. The author of The Equatorie of the Planets, for example, has been identified as Chaucer, in part because of a characteristic of his style, a high proportion of Romance words.

The problem of disputed authorship was worked upon in detail by Mosteller and Wallace,[4] in their effort to determine whether Hamilton or Madison wrote the twelve Federalist Papers. A statistical

approach was used because standard methods of historical research had not, in Mosteller and Wallace's experience, settled the issue, although an earlier attempt at using statistics by Mosteller and Williams in 1941 had also proved inconclusive when sentence length was tested as a suitable criterion for distinguishing styles. The average length of Hamilton's sentences was found to be 34.55 words, almost identical to that of Madison's, which was 34.59 words. In 1959 Mosteller received a clue to distinctive attributes of style from D'Adair who discovered that Hamilton used the word while where Madison used whilst. Since authors sometimes change their usage, Mosteller and Wallace looked for more evidence. Some was found in Hamilton's frequent use of the words upon and enough. Frequency counts of these words in the disputed papers pointed to Madison as their author.

Since Madison could have merely edited them other marker words were sought to corroborate the above finding. Hamilton was found to use the words by and from less often than Madison but to more often. Since all these were function words, it was very unlikely that the frequency counts were due to the content of the papers. Mosteller and Wallace concluded from the additional statistical evidence that Madison was the author of the disputed Federalist Papers.

Although a trial and error technique, the statistical approach has, therefore, at least as much scope as conventional literary research. In problems of disputed authorship, a literary analysis of semantic content can be a disadvantage, since each

researcher has his own bias. In contrast statistics excludes it through an objective assessment of elusive criteria.

1.1.1.2 The second branch of quantitative linguistics, the study of the optimal systems of language structure, belongs to what De Saussure[5] calls semiology, the study of the system of signs expressing ideas. Herdan[6] applies this study to the written word. In his view, Morse Code approximates an optimal coding system in that letters are systematically represented by all possible combinations of dots and dashes up to length four, the most frequent letter being assigned the shortest code, and numbers by combinations of length five. In natural language the study of the constraints in the number of possible sequences of letters (or phonemes) and in word length constitutes a part of semiology.

These coding principles are pertinent to the study of meaning, their application to which may be seen as a consequence of Martinet's[7] economy theory. In it the evolution of language is claimed to be governed by two forces, man's inertia and man's communication needs, from which two kinds of economy follow. One called syntagmatic economy consists of the reduction of the length of a word (or lexical unit) which usually expresses a frequently used concept. An example of this economy is the replacement of 'machine à laver', a long form, by Bendix, a shorter form. Paradigmatic economy, which takes place when a concept does not occur often in the language, consists of absorbing new concepts into a language without additions to the vocabulary, although at the expense of longer items in the

text. The combining of machine, à and laver into the lexical unit 'machine à laver' before the arrival of Bendix was an instance of this type of economy. These economies affect computational linguistics. The presence of syntagmatic economy makes it necessary for words (or lexical units) to be classified by an elaborate structure of semantic components in order that a mechanical intelligence may understand a text. For example, the connection between words such as chair (a case of syntagmatic economy, since it is a reduction of 'something one sits on') and sit has to be shown by components that represent the meaning of chair as 'something one sits on'.

1.1.1.3 The third branch of quantitative linguistics, the economics of mechanical translation, involves empirical examinations of the immediate environments of ambiguous forms for an approximate resolution of them. Van Buren[8] in his definition of lexical items ('multiverbal items' as he calls them) is groping along these lines. He defines multiverbal items as combinations of words at least one of which totally predicts, in certain environments, the occurrence(s) of the other word(s). For example, the word hot in 'hot dog' predicts the occurrence of dog or more precisely the specific meaning of dog (dog meaning sausage), and consequently dog is disambiguated by hot. Upon such predictability depends Booth's function number technique, which will be described in Section 1.3.2.1.1.

The application of statistical semantics to the problem of multimeaning was advocated by Weaver[9] in 1947. He envisaged not just the scanning of the words surrounding an ambiguous one, but a complete

investigation to find out which part of the context was most useful in reducing ambiguity and at what point increasing scans brought diminishing returns.

An actual investigation has been made by Kaplan.[10] With ideas similar to those of Weaver[11] he compares the effectiveness of the immediate context in reducing ambiguity with that of the whole sentence. He initially speculates that the effect of context would be most marked on homonym ambiguities where, for example, blow meaning 'to blossom' is easily distinguished from blow meaning 'to pant' and least marked where the different meanings of a word are most closely related to each other, as in the case where blow can mean 'to produce a noise by blowing', 'to pant or puff' or 'to talk loudly or boastfully'. To test his hypothesis, the following procedure was adopted: Translators were given ambiguous words, each of which was assigned a list of possible meanings, and a series of utterances in which these words appeared. The translator was instructed to select the contextual meaning of an ambiguous word for each utterance. The results of the experiment revealed the following information: the word after the ambiguous one reduces multimeaning more effectively than the word before it; two words on either side are almost as effective as a sentence; words with many meanings are as effectively reduced as those with only a few. In addition, lexical words were far more effective than function words.

For the translations of a word between which the differences in meaning are subtle, Pimsleur[12] established 'transemantic frequency

counts'. These are frequency counts of target language translations of a given source language word, which show the probability of occurrence of each of them. By means of these counts one may eliminate from the computer memory the translations least likely to occur and thereby save on machine operations, albeit at the expense of the quality of the translation. The remaining, high frequency translations, the 'cover words', would be used instead of the eliminated words to provide a coherent although stylistically unrefined translation. Thus while 'the roof is laden with snow' is the idiomatic translation of the German sentence, 'Das Dach ist schwer von Schnee', the machine would be programmed to provide the translation 'the roof is heavy with snow', nonetheless, to avoid the extra machine operations needed to decide when heavy should be used and when laden.

Reifler[13] and Mersel[14] similarly adopt the criterion of frequency of occurrence in their classification of utterances as idioms. Theoretically whole sentences could be treated as such, but their number would be infinite. The stock of idioms set up for mechanical translation will, therefore, usually consist of short phrases set up in consideration of the TL. In Reifler's example the English phrase 'the fundamental idea', corresponds to an acceptable literal German translation, 'die grundlegende Idee' and does not, therefore, have to be classed as an idiom to meet the minimum requirements of translation. However, the phrase would be so classified where the more idiomatic translation 'der Grundgedanke' is desired, namely, in the type of texts in which the phrase occurs often.

The economies provided by statistics have a place in the setting up as well as the use of the dictionary. Statistical data help one to determine how large a dictionary must be in order to contain all the words most likely to be needed for a given subject and how to arrange entries in order of frequency to reduce searching time. Parker-Rhodes'[15] statement, however, that statistics is only marginally useful in establishing procedures for mechanical translation, although of value in their application once they have been established expresses an appropriate general impression.

1.1.2 Studies of the distribution of words to place them into syntactic slots have been made for a long time. For example, through a study of the distribution patterns of sets of words like _cassage_, _cassement_ and _cassation_,[16] the suffixes _-ment_, _-age_ and _-tion_ although different in meaning may alike be categorised as nouns.

Likewise in lexicography the meaning of somewhat synonymous words such as _rompre_, _briser_ and _casser_, may be distinguished by their distribution patterns. The degree of synonymity between the words would be indicated by the similarity of the patterns, although the same criterion may apply to antonyms, too. Dubois[17] carries out a case study with the adjectives _aigu_ and _pointu_, both meaning _sharp_. In it he pinpoints the semantic categories of the nouns followed by these words and finds that among nouns admitting adjectives like _effilé_ or _arrondi_, _pointu_ occurs where _aigu_ can, but not vice versa. Among nouns admitting the use of adjectives like _sourd_ and _perçant_, _aigu_ appears where _pointu_ can, but not vice versa. The word _pointu_,

therefore, is the generic term in the first case, aimu in the second. While the distribution pattern technique does not show how the different semantic content of each word is structured, it does reveal subtleties of usage that a speaker of a language is not consciously aware of, but which a computerised semantics may ultimately have to take into account.

Methods borrowed from psychology have been used to determine a word's meaning through a listing of its paradigmatic contexts. One such method is factor analysis described by Barthes,[18] in which a word is defined by its proximity in meaning to one member of each pair of antonyms. Usually the proximity is measured on a seven point scale. Thus if visa were the word to be defined and authorisation and ban constituted one of several antonym pairs, the word's total synonymy with authorisation would be represented by a rating of one and synonymy with ban by a rating of seven. Since visa is not in fact totally synonymous with but merely more closely related to authorisation in meaning than to ban, a rating of two would probably be assigned for this pair of antonyms. Another pair might be hot and cold. Since in this case visa is equally unrelated to both antonyms, a rating of three and a half would be assigned for this pair. This rating is misleading since its point of reference is ambiguous. It is unclear whether a rating of three and a half means that visa is very much related or very much unrelated to both words. Weinreich[19] appropriately points out that factor analysis could be useful for eliciting the affect of some words if a statistical analysis were made of several people's responses, but that the method

would only incidentally reveal the core meaning of a word.

Another approach in which paradigmatic contexts are considered is that of free association.[20] In it, several people are presented with a word and are asked to state another word which it reminds them of. The responses are then sorted out to eliminate private meanings and frequency counts are made of each remaining response. Of the remaining words the one elicited most frequently is considered to be the one most related in meaning to the original word. Unlike the factor analysis approach, this one reveals hyponymy relationships, as when the specific term frigid, for example, is observed to sometimes elicit the generic term cold, but almost never vice versa.

1.2 Information Retrieval

Statistical methods are applied to some aspects of "information retrieval", a term used to describe a wide area of activity. Sharp[21] defines it as follows: "It is generally taken to embrace the whole field of the problem of recovering from recorded knowledge those particular pieces of information which may be needed at particular times for particular purposes...." Information retrieval may be subdivided into two principal areas, fact retrieval and document retrieval. This dichotomy is not always clearly recognised in the use of terminology in the literature. Document retrieval is concerned with the problems of selecting a document on a given subject area from an already classified series of documents,

and fact retrieval with the problem of summarising the content of documents to make them amenable to classification. The two areas complement each other, but present different problems.

In a fully automated document retrieval system, the actual encoding of a user's request expressed in a natural language would be performed mechanically in the retrieval of a document, as well as the matching of this code with the codes of existing documents. However, while a lot of research has been directed towards the organisation of knowledge, there has been no attempt to show how it relates to the organisation of language. Document retrieval will be treated in chapter 3 in an examination of classification schemes.

The scope of fact retrieval varies according to the researcher. A summary of a text (an abstract) by machine may be expressed in natural language or in a notation. In the encoding of a user's request two considerations are involved. One pointed out by Luhn[22] and Salton[23] is that a user may be more interested in what is original in a text than in what general subject it comes under. The other has to do with the type of user. For example, a marine biologist may need a different summary of a text on fish from a fisherman. There are two varieties of fact retrieval, derivative indexing and assignment indexing. The former is based on the principle behind the human indexer's technique of underlining important words in that a summary is derived from the words of the text itself. In assignment indexing, on the other hand, a summary is not directly formed from such words, but from a notation that

interprets them. Of the two forms of fact retrieval derivative indexing is the simpler one and, according to Coyaud and Siot-Decauville[24] has been in existence the longest.

1.2.1 Edmundson and Wyllys[25] suggest that in derivative indexing words, to which a significance rating is assigned, may be selected according to positional, semantic, or pragmatic criteria. A positional criterion would be said to apply if the first sentence of each paragraph, for example, formed part of a summary or if words in text were rated significant on the basis of their occurrence in titles, for example, where a writer might be held to choose his words with great care. A semantic criterion would be said to be employed if a semantic categorisation of words of the type summary and conclusions were utilised; here the significance rating of a word would depend on how comprehensive it was. A pragmatic approach is said to be adopted when criteria are invoked which do not directly arise from the text, such as the occurrence of the names of specialists in a field. Whichever criteria are invoked, derivative indexing tends to come within the province of quantitative statistics.

1.2.1.1 In his key-word-in-context method (KWIC) Luhn[26] attempts to base indexing on positional and statistical criteria on the hypothesis that the frequency of a word, since writers tend to repeat words as they advance their argument, and its position in a sentence are important for determining its significance. Words so graded as to their significance would constitute a pattern representative of the content of a text and texts having to do with similar topics would

possess similar patterns. Two immediate drawbacks may be pointed out. First, many words such as and and the, which occur frequently, are not useful for indexing and will have to be so designated by inclusion in an antidictionary. Secondly, writers tend to use synonyms for stylistic variation, which reduce the chances of important words appearing prominently on a frequency list.

A refinement of Luhn's statistical approach is suggested by Edmundson and Wyllys,[27] who recognise that a term that is sufficiently rare in general usage might not occur often enough to rank high in the frequency count of words, even in a text where it is important. Since according to information theory it would have a high content of information, it would therefore be important in indicating the subject matter of a text. To identify this type of word, the ratio between a word's frequency in a specific text and its frequency in general might be examined.

For the above mentioned type of term, "special sets of reference frequencies for special fields of interest"[28] would be kept. For each field a vocabulary of words is compiled and the frequency of occurrence of each word in a statistical sample of texts that belong to the field is calculated. In addition, the total number of words in the sample is counted so that the percentage of words that each word of the vocabulary represents may be calculated. By calculating the percentages for words in a particular passage one may determine its field. Certain percentages for words such as trauma, sensory and behavioural, for example, would indicate that

a given passage belonged to the field of psychology. If the percentages for a minority of other words in the passage failed to fit the pattern for psychology, this fact would be taken as an indication that another field was involved. The frequent occurrence of the word chromosome, for example, in a passage belonging to the field of psychology might suggest that it was about the hereditary factor in human psychological makeup. The above refinement in statistical procedure is the making of the notional family[29] and idioglossary approaches, which will be discussed in section 1.2.2.1 and 1.3.2.1.3.

1.2.1.2 In Harris'[30] string analysis positional criteria are used for indexing. Each sentence is analysed into a formal centre and the right and left adjuncts. Words in the formal centre are considered significant and those in the adjuncts redundant. In the sentence 'Today automatic trucks from the factory which we just visited carry coal up the sharp incline', the formal centre, 'trucks carry coal', would form the extract. Other examples, however, support Coyaud's contention that the formal centre does not always contain the most important information. In Noel's[31] sentence, 'Additional information concerns availability of microfilm services' the main topic is found in the phrase, 'microfilm services', which is an adjunct.

The 'Sentence Dictionary' of Earl and Robison[32] like string analysis is based upon the hypothesis that topic sentences may be identified by their structure. To classify indexable sentences, a large sample from a total of nine chapters out of books selected at random from the Palo Alto library was sorted according to structural

type. In the first analysis a sentence was counted as a sequence of parts of speech, but the 3098 types of sentence discovered were found to be too high for computation. Subsequently sentences were counted as sequences of phrases to reduce the number of types. The topic sentences found in the books were in addition placed in an index. Since at the time of writing the dictionary was still in the experimental stage, it is not known whether or not the topic sentences in their structure form a distinct grouping. However, even if the "Sentence Dictionary" is only partially successful it will still be of value to derivative indexing.

1.2.2 In assignment indexing the notation (or documentary language, as Coyaud and Siot-Decauville[33] call it) expresses the relationship between synonymous utterances for application in what Salton[34] calls language normalisation programmes. Of those that operate on sentences there are two well-known types. In one, the aim is to reduce complex syntactic constructions to a group of equivalent simple kernel sentences with a specified canonical pattern such as the noun and verb one. Rigorous rules, however, have not been formulated to carry out the aim. The other type is the transformational approach, in which surface structures such as 'the man eats the food' and 'the food is eaten by the man', are recognised as equivalent through an analysis of the active and passive voice. Below the sentence level the thesaurus approach, in which words in a text are replaced by corresponding thesaurus heads, is a form of language normalisation in which synonyms are eliminated and redundant words are ignored, although information is lost in the type of

thesaurus in which a specific term is replaced by a generic one.

1.2.2.1 Luhn[35] advocates a thesaurus approach based on statistical criteria to colligate the various levels of specificity on which authors express similar ideas. Such a thesaurus would consist of 'notional families' (groups of words related in meaning) compiled by experts in the field from which the texts to be indexed are drawn. Each word in a text would be assigned a keyword or a concept number according to the family to which it belongs. The existence of a notion would depend on its likely frequency of use. Since in the field of electricity, for example, the words subsumed under the notional category electricity would predominate about equally in most of the texts, the words would be partitioned into more specific categories. At the other extreme the notion butterfly in texts on electricity would probably appear too rarely to discriminate between texts, so that a more generic notion like insects, or even 'living things' would be more appropriate.

The thesaurus having been compiled, the words in a passage are analysed and frequency counts are made of the notions. The most frequent ones, which are considered to be the most representative of its content, form a 'mechanically prepared notional abstract', an encoded summary. A refinement of the notional family approach might include a supplementary index by which to regroup words under different notions according to context. If the word butterfly occurred frequently, for example, in a passage on electronic butterflies and on the basis of its usual frequency the notional

category insects had been set up, the computer might be programmed to replace it with the category butterfly. The set of notional families would thus constitute a general classification with criteria for pigeon-holing texts by machine.

1.2.2.2 In the interlingual approach to language of the Cambridge Laboratory Research Unit thesaurus heads having a wider scope than Luhn's notional categories were set up with a view to making mechanical translation, library (document) retrieval and mechanical abstracting amenable to the same treatment. Masterman, Needham and Sparck-Jones[36] claim that "the very nature of the problem of interlingual mechanical translation is like that of information retrieval in that it demands a general, that is, a logical approach". The "logical approach" consists of linking the various surface structures of languages to a common deep structure, which constitutes an interlingua.

One function of the thesaurus is to resolve multimeaning. Accordingly, concept numbers, which represent heads in Roget's Thesaurus, are set up and words are listed under them, ambiguous words being placed under several concept numbers. For the word plant in Masterman's[37] example there are three concept numbers 184, 300 and 367, depending on whether it means to place, to insert or a vegetable respectively in a given context. If the context includes another ambiguous word flowering, for example, which may be found under the numbers 5, 161 and 367 depending on whether it means essence, produce or vegetable respectively, the concept numbers for both words would be

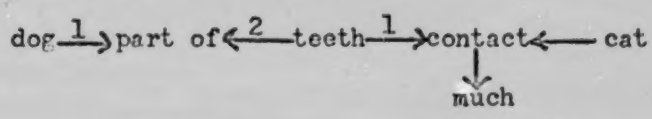
matched against one another. Since the number 367 is possessed by both words, it is accepted as representing their contextual meanings. In providing explicit indices by which to make disambiguation computational the thesaurus approach of C.L.R.U. is the starting point for a comprehensive computerised dictionary, the construction of which will be discussed in chapter 4.

The thesaurus heads represent not only words but also stems and endings ("chunks" in their terminology). A chunk is "the smallest significant language-unit which can exist in more than one context, and which for practical purposes, it pays to insert as an entry by itself in an MT dictionary". The Italian word piantatore, for example, consists of three as follows: piant, at and ore. In a monolingual thesaurus a chunk may be represented in the form of a tree and when it is connected with a tree of another language for translation, "the two trees together form a lattice each point of which looks both ways and is itself a translation point".

The greatest challenge to the interlingual approach is the representation of syntax. Parker-Rhodes[38] has found that because part of the meaning of a sentence is conveyed by the choice of words (lexically) and part by the manner of their combination (syntactically), to a different extent in different languages it is difficult to make translation computational. To relieve the difficulties of word order he advocates the use of affixes to replace syntactic structure for conveying information. For example, in the phrase 'race horse' the role of the word order, which contrasts it with

'horse race' would be taken over by a role-indicating affix added to one of the words. Whether the representation is 'race-R horse' or 'horse race-R' would depend upon the word order of the target language.

Richens[39] represents syntagmatic relations by means of a "semantic net" that links concepts ("naked ideas") that are not specific to any individual language. The sentence, 'the dog bites the cat', for example, is represented by the following two-dimensional arrow structure:



The concepts are designated by words simply as a convenience to the observer. The diagram is essentially an explicit paraphrase of the original sentence inasmuch as it could also be said to represent the sentence, 'the dog's teeth have much contact with the cat'. The dependency links indicated by the arrows appeared later in Schank's[40] research. At the time of writing Richens had found no general mechanical procedure for extracting semantic nets from a text.

Parker-Rhodes'[41] "interlingual formulae" resemble Richens' "naked ideas" except for the format of binary bracketing, which determines the surface structure that the formulae will take in a given language. For the clause 'dexterity which cheats the eye' for example, the formulae are as follows: ((eye cheat) type) (hand skill)). By rearranging the brackets according to given rules of interlingual grammar one may represent various paraphrases of the

clause. According to one rule, which centres on the word type (a "weak element"), the formulae ((eye cheat) type) may be transformed into the synonymous construction (eye (cheat type)), which represents the phrase 'eye-cheating dexterity'. According to another rule, a redistribution of thesaurus heads allows for the paraphrase 'visually deceptive'. In order to implement these rules a computerised lexicon (or thesaurus) is necessary to show how the formulae are created by a step by step collation of the dictionary entries for each word.

1.3 Mechanical Translation without Interlingua

While the C.L.R.U. regarded translation as a two-way process involving the representation of a source language by an interlingua which provides output in a target language, specialists in mechanical translation alone have looked at it as a one-step process, in which the source language is translated directly into a target language. Various nonlinguistic models have been suggested for this one-step process.

1.3.1 One is Weaver's[42] cryptography analogy, which is based upon the observation that by making frequency counts of letters and combinations of letters for a given language, one can decode a message written in it. In a letter to Norbert Wiener (1947, March 4th.), he suggests treating cryptography and mechanical translation analogously so that a text in Russian, for example, would be visualised as an English one coded in strange symbols. The value of the analogy is limited, since only at the sentence level is a one-to-one correspondence

likely to occur between the utterances of two different languages, and the number of possible sentences in a language is infinite.

Another model, adopted by Nida,[43] Jakobson[44] and Yngve[45] independently, is based on communication theory. A few quotations will throw light on the value of the model. In making the point that translation is not word-for-word but thought-for-thought Jakobson[46] says "...translation from one language into another substitutes messages in one language not for separate code-units but for entire messages in some other language". In a more elaborate use of communication theory terms, Yngve[47] claims that "The function of the message source is to select the message from among the ensemble of possible messages. The function of the rules of the code or codebook is to supply the constraints of the code to which the encoded message must conform.....The function of the decoder is to recognise the features of the encoded message that represent the constraints of the code, remove them, and supply the destination with a message that is a recognisable representation of the original message". In this passage the use of the phrase "rules of the code or codebook", for example, instead of 'grammatical rules' serves to emphasise that natural and artificial languages may be analysed by the same linguistic methods. This recognition of parallels between linguistics and communication theory is of cross-disciplinary interest. However, it does not provide any insights into mechanical translation.

Ceccato's[48] model concerns the mentalistic processes behind linguistic performance. According to him, translation is possible

only if, by semantic connections, the SL text can be replaced by the thought that it represents, which, in turn, can be replaced by the TL text. The model shows the step by step construing of dictionary entries by which a human understands an utterance. Upon finding that the first word in a sentence is the or tree, the translator sets up in his mind a list of the types of words that may follow, a "correlational structure". Upon finding the second word he links it with the first to form a "correlational net". As additional words in the sentence are fitted into the net, the structure of it is adjusted accordingly. For the sentence 'John she loves' of dubious grammaticality the translator tests the binary connections, 'John loves' '...loves John' and 'she loves' and by a process of elimination selects the second two as the valid ones. An examination of the pairs reveals John to be the object, loves the verb and she the subject of the sentence and a correlational net is formed accordingly, which Ceccato illustrates with squares containing dots. The dots in the lower squares represent the words, loves and John, the dot in the upper square, she. The replacement of dots and squares by lines would make it apparent that correlations are in fact immediate constituent analyses. In Ceccato's model new terminology is applied to old techniques.

1.3.2 The first empirical linguistic attempt to grapple with the problems of MT was the Georgetwon-IBM experiment carried out by Dostert[49] and others in translating English into Russian. MT was divided into two operations, one of selection, in which lexical data is handled to produce the correct TL word, and one of manipulation,

under which word order was subsumed. The penetration of the problems was not deep, since the experiment was on a small scale. Although dependence upon post- and pre-editing was eliminated, the dictionary consisted of less than 250 terms and no more than two English equivalents were assigned to each ambiguous word.

The programme dealing with the dictionary component was divided into five operations. The first covered SL and TL words that were in one-to-one correspondence with each other. The second treated multimeaning problems that could be solved by examining the word before the ambiguous one. The third handled those that an examination of the word after the ambiguous one could solve. In the fourth operation words in SL that were superfluous in TL were omitted. In the fifth, terms missing in SL that were required in TL were added.

In this early experiment, then, the criteria for translating were based on the physical rather than on the structural context of utterances. This distinction may be observed in an analysis of the sentences, 'I painted the white wall' and 'I painted the wall white'. A translation by structural context would take into account the difference in IC structure between these sentences and Schank[50] and Weinreich[51] (chapter 2, 2.2.2 and 2.2.3.2) would, in addition, relate it to the organisation of non-linguistic knowledge. But a translation by physical context would merely take into account the difference in the word order of wall and white.

1.3.2.1 Mechanical translation in the 1950s, following the Georgetown experiment, continued to be based on physical context.

Further research produced a miscellany of ambiguity types, - predicate block structure and inflectional ambiguities, homographs, orthographic coincidences, and contextual and punctuation problems.

1.3.2.1.1 The first two types of ambiguity[52] concern words that have one meaning but many possible translations. Ambiguities in a predicate block structure are said to occur when a word has one meaning but many possible translations into the target language, due to the syntactic relations into which it may enter. The Russian word sdelano, for example, is such an ambiguity as it may be translated into English as done, is done or as be done depending on whether it occurs with an auxiliary verb, with no subject or auxiliary or with pust' respectively. Since is and be are function words, they may alternatively be relegated to syntactic analysis so that done remains as the translation of sdelano. Ambiguities in predicate block structure are only classifiable as ambiguities because syntactic analysis was based in the 1950's on the physical and not the structural configurations of words.

Inflectional ambiguity has to do with morphology and is said to occur when the number, gender or case of a word is not clear. In the Russian example of Janiotis and Josselson[53] the word stanchii is such an ambiguity because it may be genitive, dative, locative singular, nominative or accusative plural. Inflectional ambiguity covers Reifler's[54] distinction between monogenetic and polygenetic meaning. In his German example, the word aus (out of) is said to have monogenetic meaning because it can take one case only, the dative.

The word dieser (this) has polygenetic meaning since this form may have three meanings depending on whether it is singular and masculine nominative or a feminine genitive or dative or is a genitive plural. From this example it may be observed that "inflectional ambiguity" is synonymous with "polygenetic meaning."

Homographs and orthographic coincidences[55] are words that have many unrelated meanings. The latter type of ambiguity in addition covers such words that belong to the same grammatical class. The Russian verb plachu (I weep or I pay), which is inflected from both plakat' (weep) and platit' (pay), is an orthographic coincidence. On the other hand dam, which may be either the first person singular of shdat or the genitive plural of dama is simply a homograph. The ambiguity types mentioned so far do not appear to have been formulated according to the types of computation involved. In such a formulation predicate block structure, inflectional ambiguity and homographs would be categorised as types of ambiguity that can be resolved by parsing. Orthographic coincidences would be grouped with contextual problems as types that cannot be so resolved.

Contextual problems[56] are types of ambiguity that make a computerised semantics necessary. An example of such a problem is the word board, which may mean piece of wood, food (as in 'room and board'), stage, council (as in 'board of directors') or an action as in 'to board a train'. While this last meaning can be identified because the word is a verb, board in all its other meanings is a noun. In some cases a whole sentence will not solve this type of problem.

In Lukjanow's[57] example, the Russian sentence 'Ia Zaplatila za stol' ('I paid for the table/board') stol can be translated as table or board. The limitation of disambiguation to within one sentence, may have been convenient for early attempts at mechanical translation and within the tolerable limits of inaccuracy according to Kaplan's study but it does not have a linguistic basis.

The adoption of one pragmatic criterion in disambiguation, that of limiting the context under examination to within one word, allows direct syntactic links to be utilised. Such links are represented by concept numbers in Booth, Brandwood and Cleave's[58] method. This approach is similar to the one used by Masterman, (section 1.2.2.2), except that in hers ambiguous words are scattered among several concept numbers and are looked up by means of an alphabetised cross-reference dictionary, while in Booth's method the different numbers representing a word's meaning are listed together. This technique was applied mainly to prepositions, which occur frequently. One such preposition is the German word, auf, which may appear in the phrases, 'auf dem Tisch', 'auf dem Tanz' and 'auf dem Lande' meaning respectively 'on the table', 'at the dance' and 'in the country'. In Booth's notation the various possible translations of "auf" are represented as follows: ("auf"=1) on, 2)at, 3)in) and correspondingly the German nouns as follows: ("Tisch"=1)), ("Tanz"=2)), ("Lande"=3)). The matching of numbers provides the translation. In the case of 'auf dem Tisch', the concept number for Tisch is found to be identical to the number for auf meaning on, so that the translation, 'on the table' is given. By the application of such

numbers on a larger scale, the correct preposition may be supplied for every noun in English.

The representation of syntagmatic links by means of concept numbers is a means of detecting idioms. So applied, such numbers are called by Booth, Brandwood and Cleave[59] function numbers. These uniquely represent the words that can be part of a particular idiom. When a word with such a number is detected, the words following it in the text are tested for possession of the same number. For example, the words, il, y and a, in the French idiom 'il-y-a' might each be assigned function number, 1. Upon finding il in a text the computer searches the words following il. If y and a follow, the idiom translation, 'there is' is supplied. Otherwise a literal translation is assigned by default.

Pragmatic solutions to translation problems include the manipulation of punctuation. In German, capitalisation is usually an explicit criterion for disambiguation, since it is applied not only to words at the beginning of a sentence but also specifically to nouns. The comparative dichter (tighter) thus differs in form from the noun Dichter (poet). This distinction does not apply, however, at the beginning of a sentence. In 'Dichter ist der Hahn (faucet/cock) geworden' ('The faucet/cock has become tighter/ a poet') Dichter contributes to the ambiguity of the sentence. To receive the full benefit of the German convention Reifler[60] advocates the reservation of capitalisation for nouns exclusively so that the first word of this sentence would be dichter from which the machine would derive the translation tighter.

Various punctuation problems might disappear through manipulation. In French,[61] the apostrophe and the dash are ambiguous in that they sometimes separate two different words and sometimes two parts of the same word. An apostrophe divides the single word aujourd'hui and the two words l'or and similarly a dash, the single word porte-clé and the two words vient-il. The rejection of the convention that so separates two distinct words would resolve the problem. However, while such a manipulation was a convenient stopgap measure in the 1950's, it is no substitute now for effective procedure. Parts of a word might be distinguished from complete words by Booth's function numbers.

1.3.2.1.2 The stems and endings method, which Booth[62] and Richens first applied in 1947, was a way of segmenting phrasal compounds for economy in the inventory of items in the lexicon. A group of many vocabulary items like seaboard, seaside, seaway, board, way, boards and ways would be atomised into fewer forms, sea-, -s, -board, -side and -way with increasing economy as more and more words are partitioned. This approach may be applied bilingually too. In the German words, Musik and Direktor,'k' would be segmented from the rest of the word to implement the rule that 'k' becomes 'c' in a translation into English. Hybrid compounds like Goldhandel (gold trade), however, would not be amenable to the same treatment.

Against the economy in the inventory of elements in the lexicon, expecially in an inflected language like Russian,[63] the drawback of the additional complexity to the grammar of a language

necessary for generating words from stems and endings must be balanced. Reifler[64] indicated that it was on this account that he did not adopt the method. At the time of his criticism, however, he had access to photoscopic disc, an improvement in technology that enabled a computer to absorb a relatively large vocabulary.

A part of the complexity of grammar following from the stems and endings technique lies in the careful setting up of them so that one word may be partitioned by computer in as few ways as possible. The opportunity to control the setting up of stems and endings exists, when letters or letter sequences can be part of either and thereby constitute what Reifler[65] calls an "X-factor". The Russian word, r'iuopovu (fisherman), contains one. The usual dissection of this word is r'iu-o-povu, where 'o' constitutes a connector and povu means 'to the catcher'. Since the existence of the free forms, r'iu ('of fishes') and opovu ('to the tin'), makes the incorrect translation, 'to the tin of fishes', possible r'iuopovu, is divided for the purposes of translation into r'iuo and povu instead of into r'iu and opovu. The connector, 'o' is the X-factor since it is the crucial element in avoiding incorrect construing.

For some words the number of possible partitions into stems and endings cannot be reduced because of inherent ambiguity. For example[66], the German word Wachtraum, may be split into either Wacht and Raum (guardroom) or Wach and Traum ('waking dream'), the 't' being attachable to both stem and ending. Since the ambiguity lies in Wachtraum itself, partitioning must be based upon a resolution of the

word's meaning in context.

1.3.2.1.3 One of the first pragmatic attempts at resolving ambiguous words in a text consisted of a categorisation of the various meanings of each word according to subject area. A special dictionary containing words so categorised was called an idioglossary (or micro-glossary). Which idioglossary to apply to a text was determined by indexing it either by machine or by a pre-editor. When it was first introduced, it was a stopgap measure to prevent a computer's very limited memory space from being wasted upon words not applicable to the types of texts to which mechanical translation was applied. Dostert[67] probably had the concept of the idioglossary in mind in 1955, when he suggested that a "functional lexicon" be used "...when a text in a given functional field area is being translated". The word stream, for example, would be entered into two such areas, one consisting of geographical terms and the other of engineering terms, disambiguation then being dependent on the content of the text under consideration.

The content of the text as a whole is determined by the type of frequency counts of words made by Luhn, p. 12. In addition subdivisions may be recognised so that where a text fits into two subject areas two frequency counts may be made, one for the local context of an ambiguous word and one for the whole text. Such a procedure might be useful if a text had to do with the social implications of atomic energy, for example.

The task of structuring a system of idioglossaries was

undertaken by Micklesen.[68] Like Luhn's notional categories, his system was arrived at intuitively with adjustments made through the observation of statistical data. The notation for representing the idioglossaries was decimal. Digits in the tens' column were reserved for the major divisions of knowledge and those in the units' column, for their subdivisions. A word considered to belong to mathematics in general, for example, would be assigned a number such as 10. A term belonging to a particular branch of mathematics would be specified by the replacement of the digit zero. Thus the number 11 would designate an algebraic term and 12, a geometrical one. This type of notation was not originated by Micklesen, but was in fact a variation of Dewey's decimal classification. Whereas Dewey applied it to retrieve documents, Micklesen designed his system to categorise words.

Since Micklesen did not have access to a computer the words he had categorised were checked in the manner of a machine against actual texts to test the validity of his idioglossary system. He found that 88% of the words were correctly assigned. These results serve to emphasise the complexity of the organisation of knowledge with which a computerised semantics must come to terms.

1.3.2.2 While the emphasis in mechanical translation in the 1950s was on the use of the lexicon, its limitations were recognised. While Perry's[69] experiments revealed that, a translation without a grammar, when applied to scientific and technical material, was comprehensible, members of the MIT school, including Bar-Hillel and Yngve[70] advocated

the systematic use of parsing to obviate the necessity for a proliferation of ad hoc rules, like the one stating that if German der follows a capitalised word with no intervening comma, 'of the' will be the translation 95% of the time. The advantage of a parsing programme was that rules set up for disambiguating a given word are equally applicable to other words that belong to the same paradigm. The method of parsing valid for the article der, which may be nominative, genitive or dative, is equally valid for the words, dieser (this) and jeder (each) whereas ad hoc rules apply only to individual words.

Yngve's tenet that syntax should be handled before selectional restrictions anticipates the main drawback of Katz and Fodor's[71] marker theory, a major development in the 1960's in making semantics computational. Yngve says: "The selectional relations between words in open classes, i.e. nouns, verbs, adjectives and adverbs...can be utilised by assigning the words to various meaning categories in such a way that when two or more of these words occur in syntactic relationships in the text, the correct meanings can be selected".[72] Before the meaning of the word plant, for example, can be determined by that of flowering, it must first be ascertained that plant is a noun and flowering an adjective. In order to represent the semantic content of a word in a form useful for computation, the word's syntagmatic structure must be indicated.

An attempt to formally represent syntactic relations was made by Bar-Hillel[73] in his categorical grammar. The goal was to ensure first that all grammatical constructions were assigned the same

notation to contrast with that of ungrammatical ones and secondly that an utterance belonging to a given part of speech and a word belonging to the same one be identically encoded. The principles of Bar-Hillel's notational system were based upon the rules of arithmetic applied in the multiplication of vulgar fractions. The 's' and 'n' combinations, of which the formulae representing the parts of speech consisted, were set up in the form of denominators and numerators. The symbol 'n' designated nouns, 'n/n', adjectives and 's/n', verbs. The formula for a whole utterance is derived from a step by step construing of the formulae of its parts. In the sentence 'Poor (n/n) John (n) works (s/n)' the reduction of 'n/n . n' by 'cancelling out' to n represents the linking of Poor, an adjective, with John, a noun, to form a noun phrase 'Poor John'. The connection of this phrase in turn with works is represented by the reduction of the 'n' (for the noun phrase) and 's/n' combination to 's'. This symbol designates the sentence as grammatical. An utterance of the type 'Poor (n/n) works (s/n)', for example, would be reduced to 's/n', which indicates an ungrammatical sentence. Similarly 'works (s/n) John (n) poor (n/n)', which is analysed to be 's . n/n', is so indicated.

The main drawback of the categorical grammar is that of scale. By testing sentences with a transformational grammar, it may be verified that the complexity of language is beyond what the conventional categories such as noun, verb and adjective represent. Personal experience reveals that the 's' and 'n' notation with its arithmetical framework is overpowered by the demands of various types of sentence construction. However, when stripped of the procrustean

framework, the grammar suggests an atomisation of the traditional parts of speech into elements more useful for computation.

The attempts in the 1940's and 1950's to circumvent the task of organising the semantic content of a word into a computerised dictionary only succeed at all in quantitative linguistics. For mechanical translation and information retrieval, the ultimate goal is the resolution of the amphibology 'I shot the man with a gun' in the sentence 'I shot the man with a gun, but if the man had had a gun too, he would have shot me first'.[74] Resolution requires the recognition by language normalisation that 'if the man had had a gun too' implies the man did not have a gun so that 'with the gun' is observed to link with I and not man. A shorter range goal is the resolution of ambiguity in a single word and paraphrase recognition relatable to single words.

Footnotes to Chapter 1

1. Herdan 1956
2. Herdan 1956
3. Bally and Sechehaye 1959
4. Herdan 1956
5. Bally and Sechehaye 1959
6. Herdan 1956
7. Martinet 1964
8. Van Buren 1968
9. Weaver 1947
10. Kaplan 1955
11. Weaver 1947
12. Pimsleur 1957
13. Reifler 1961a
14. Mersel 1961
15. Parker-Rhodes 1958
16. Dubois 1964
17. Dubois 1964
18. Barthes 1968
19. Weinreich 1958
20. Doese 1965
21. Sharp 1965
22. Schultz 1963
23. Salton 1961
24. Coyaud and Siot-Decauville 1967
25. Edmundson and Wyllys 1961

26. Schultz 1968

27. Edmundson and Wyllys 1961

28. Edmundson and Wyllys 1961

29. Schultz 1968

30. Harris 1962

31. Noel 1968

32. Earl and Robison 1970

33. Coyaud and Siot-Decauville 1967

34. Salton 1961

35. Schultz 1968

36. Masterman, Needham and Sparck-Jones 1959

37. Masterman 1956

38. Parker-Rhodes 1961

39. Richens 1956

40. Schank 1969

41. Parker-Rhodes 1961

42. Weaver 1947

43. Nida 1963

44. Jakobson 1959

45. Yngve 1955b

46. Jakobson 1959

47. Yngve 1955b

48. Ceccato 1960

49. Dostert 1955

50. Schank 1969

51. Weinreich 1966

52. Josselson and Janiotis 1961

53. Josselson and Janiotis 1961

54. Reifler 1955

55. Josselson and Janiotis 1961

56. Josselson and Janiotis 1961

57. Lukjanow 1961b

58. Booth, Brandwood and Cleave 1958

59. Booth, Brandwood and Cleave 1958

60. Reifler 1961

61. Muller 1968

62. Booth, Brandwood and Cleave 1958

63. Booth, Brandwood and Cleave 1958

64. Reifler 1961b

65. Reifler 1955

66. Booth, Brandwood and Cleave 1958

67. Dostert 1955

68. Micklesen 1961

69. Perry 1955

70. Yngve 1956

71. Katz and Fodor 1963

72. Yngve 1957

73. Bar-Hillel 1953

74. Katz and Fodor 1963

2 STRANDS OF SEMANTIC THEORY, 1960-1972

2.1 The Notation and Functioning of a Computerised Dictionary

Linguistic research in the 1960's has explicated the difficulties to be faced in computerised semantics, but has provided no model that overcomes them. Katz and Fodor[1] attempted to create one, but theirs barely suffices to disambiguate words inasmuch as it fails to take into account their syntactic contexts. However, as a semantic theory Katz and Fodor's model may be considered the nucleus of research into computerised lexicography. In the realm of syntax, investigation mostly centres on transformational grammar. For a semantic model that covers the problems brought up by the linguists, one must turn to the type of artificial intelligence developed by Schank[2] and others at Stanford.

2.1.1 Katz and Fodor[3] envisaged the components of their dictionary as concepts independent of the operation of natural language. A full discussion of the constraints to which a language, natural or documentary, is subject will be provided in chapter 3. At this point, it may be said that the components by being called concepts do not escape reference in terms of natural language. In fact, subsequent discussion will reveal that they function syntagmatically as adjectives and paradigmatically as antonyms. As a prelude, therefore, to a consideration of Katz and Fodor's dictionary, it would be appropriate to examine antonym and distinctive feature analyses, upon which the setting up of it depends.

2.1.1.1 The beginnings of distinctive feature analysis may be traced back to the Cours de Linguistique Générale prepared in 1913 from the notes of De Saussure by his students.[4] In it the valeur of a word is claimed to be derived from its association with other words, that of the English word sheep, for example, being different from that of the French word mouton, because it contrasts with another word, namely, mutton, which refers to a live animal. Since 1913, the value of the English word has changed. Three words sheep, mutton and mouton now correspond to the French word. Accordingly the relationship between them may be stated formally by attaching to sheep, mutton, mouton and mouton the respective sets of distinctive features, /+live, +ovine/, /-live, +meat, +ovine/, /-live, +skin, +ovine/ and /±live, +ovine/. The analysis may be extended to other words such as pig, pork, cow and beef, which may be assigned the respective groups of features, /+live, +swine/, /-live, +swine/, /+live, +bovine/, and /-live, +bovine/. Dictionary entries thus encoded offer explicit indices for computational analysis. The effectiveness of such distinctive features will depend upon each of them's being assigned a unique meaning.

In being a form of language normalisation distinctive feature analysis will only incidentally represent words with the same categories as traditional grammar. In Prieto's[5] example, the sentences 'Elle le regarde', and 'Elle la regarde', le and la are respectively assigned the groups of features /+singular, +definite, 3rd person, +masculine or neuter/ and /+singular, +definite, +3rd person, -masculine or neuter/. As they differ with respect to a single feature, gender, in the above contexts, they constitute what

Prieto calls "noèmes". Since in the sentences 'Elle regarde le cahier' and 'Elle regarde la porte' the use of the wrong gender of article can be detected and corrected with absolute certainty, unlike in the first two sentences, the gender distinction is redundant. Correspondingly the sets of distinctive features for le and la are identical, both being /+singular/ /+definite/ and /+3rd person/. The function of these articles is analogous to that of the two phonemes /n/ and /ng/ in English. While in most environments they are distinctive, before the phoneme /K/ in such words as income they are not. In phonology what precedes /K/ is called an "archiphoneme". Analogously the prefix archi is applicable to the articles le and la, which thereby constitute an "archinoème".

In the examples of distinctive feature analyses presented above the convention of plus and minus signs has been adopted to reveal the explicit indices with which computational procedure has to deal. In theoretical linguistics they are often not displayed but left to human imagination. In Prieto's[6] actual example the notation did not consist of forms of the type /+masculine/ and /-masculine/ but of the type /masculine/ and /feminine/. Similarly Katz and Fodor[7] contrast /animate/ with /inanimate/ rather than /+animate/ with /-animate/. For computation either a special table of antonyms or the plus and minus convention is necessary. Since the latter would be less complex to programme a computer with, it will be applied throughout the rest of this chapter. The plus and minus signs will be called 'indicators' and that which follows them will be called 'descriptives'.

2.1.1.2 Antonyms may be divided into two categories.[8] One group which may be described as "non-gradable" covers antonyms like married and single, which do not admit of degree. The second, describable as "gradable", embraces antonyms like big and small, which do admit of degree. These resemble conversives in that when one is replaced by the other in a sentence in conjunction with a transformational rule, a paraphrase is produced. Because of the conversive relationship between buy and sell, for example, the paraphrase 'Fred bought something from John' may be derived from 'John sold something to Fred' by inverting the relative sequence of the nouns. By a similar inversion, 'Fred is smaller than John' may be derived from 'John is bigger than Fred'.

Gradable antonyms are responsible for what Weinreich[9] calls "impure linking", a type of syntagmatic relationship in which it is not anomalous for one noun to be qualified by two adjectives which are antonyms. In the sentence 'A small elephant is big' the two adjectives are not incompatible, since the word small refers to elephant standards and big to other standards. Because big and small have this property, it would be difficult to encode them within the plus and minus convention.

Antonymy along with paradigmatic relationships in general operates not so much between words as between given meanings of words. As the French word libre, for example, has many different meanings, so it has many antonyms as follows: prisonnier, captif, esclave, forcé, occupé, gané and embarrassé. In English the word animal is

sometimes the opposite of human and sometimes includes humans when it is the opposite of plant. These are what Duchacek[10] calls partial antonyms. The antonym relationship between some pairs of words applies only in certain idioms. For example, tort is the antonym of raison only insofar as 'avoir raison' and 'donner raison' are antonyms of 'avoir tort' and 'donner tort' respectively. These are called phraseological antonyms.

2.1.2 In Katz and Fodor's[11] marker theory distinctive features are organised not only into antonym pairs, but also into hierarchies. While the number of features included in their representation of the French word canard would probably not be adequate for computation in an actual experiment, it will be adopted in this discussion to explain their theory. In the Larousse dictionary[12] the different meanings (along with the English translations of them) of this word are as follows: m. ZOOL. duck; canard mâle, drake; canard sauvage, wild duck. II FAM. nag, jade (cheval); squawk (false note); hoax, false report (ou) news, canard (fausse nouvelle); rag (journal); lump of sugar dipped in brandy or coffee (sucre); marcher comme un canard, to waddle. (V. DANDINER [SE].) In Katz and Fodor's tree a selection of them is organised as in figure 1. In this diagram round brackets represent distinctive features (or "markers" as they are called) and square ones, "distinguishers", which denote that part of a word's semantic content that is allegedly not necessary for computation. This issue will be taken up in chapter 4.

For computational analysis the above type of tree would be

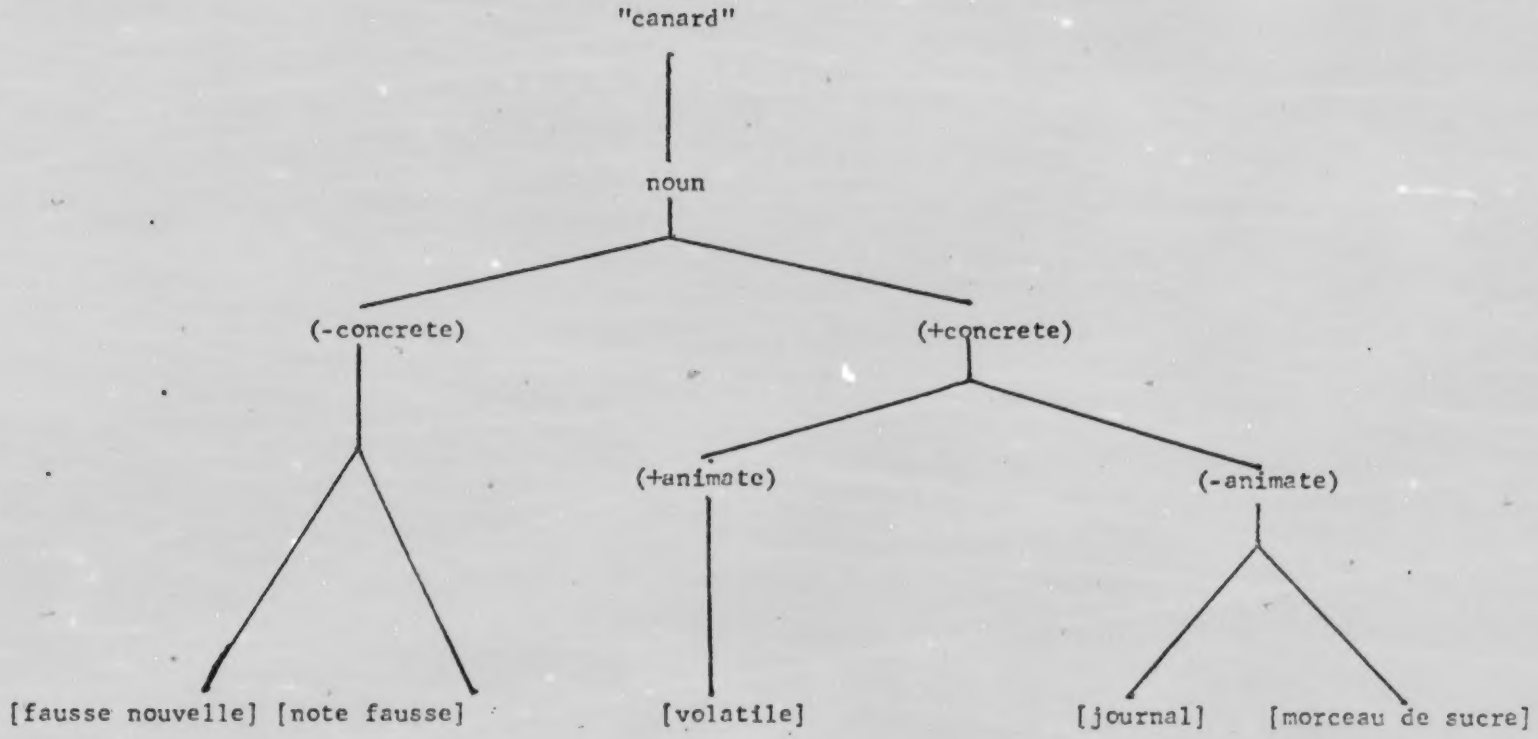


Figure 1: Marker tree.

replaced by a one-dimensional formula as follows: "canard" = -concrete (-sound [fausse nouvelle] or +sound [note fausse]) or +concrete (-animate (+square [morceau de sucre] or -square [journal]) or +animate [volatile]). In this formula square brackets have the same significance as they have in the above tree. The round brackets do not, but are in fact isomorphic with the branches of a tree.

An example, which Katz and Fodor would probably accept, of how to select the correct contextual meaning of canard may be seen in the analysis of the sentence 'le canard respire', for which an appropriate marker formula for the word respire is "Respire" = +concrete (+animate [vivre]). To eliminate the non-contextual meanings of canard the markers of each word are matched against each other. The first one in the formula for respire, which is /+concrete/, is matched against the first one in that for canard, /-concrete/. Since they differ, the contents of the enclosed round brackets beyond /-concrete/ are ignored and analysis starts again after the second or. Since the markers for both words this time are identical, being /+concrete/, the second marker for respire, /+animate/, is located and searching is now limited in canard to the confines of the bracketed portion following /+concrete/. Since the marker for respire does not match /-animate/ for canard, the markers within round brackets that follow this one are passed over and analysis proceeds after or, whereupon a match is found. Since there are no further markers for either word, marker analysis ends and the distinguisher [volatile] determined by the match, is extracted as an indication of the contextual meaning of canard. That the requisite

distinguisher is found at the end of the formula in this example is purely coincidental. With a different arrangement of markers it might have been located in the middle.

2.1.3 While from the discussion so far the utility of Katz and Fodor's theory appears to depend upon its capacity to resolve ambiguity, other claims have been made for it. Postal[13] suggests that "the semantic component provides each sentence with a semantic interpretation in the form of a set of readings and accounts for the speaker's knowledge of the facts of meaning." Nida[14] claims that Katz and Fodor's tree could handle the distinction between the central and peripheral meanings of a word, through the location of the former on the left branches and of the latter on the right. These two claims seem to be based on the appearance of the tree rather than on its actual functioning in computation.

In light of Bolinger's[15] criticism of Katz and Fodor's theory, it appears that the small scale on which they envisaged setting up trees would be insufficient. While it is extreme to claim that the necessity to add markers to one of Katz and Fodor's trees as it was tested on sample sentences invalidated their theory, his findings suggest that the content of marker trees will have to represent not dictionary but encyclopedia entries in order to be functional.

Because of the nature of encyclopedic knowledge the representation of it requires flexibility in the structure of the marker tree. While Katz and Fodor found a set of markers amenable to

hierarchical organisation for the word canard, such a set is hard to find for the English noun trunk. While a comprehensive notation for the word cannot be determined within the scope of this discussion, since it would require a lot of empirical data, a suitable formula for the meanings of trunk as a noun would appear to be of the following type: "trunk"= +container,...[box]; -timber [elephant's nose] or +timber [portion of tree]. In this formula the departure from Katz and Fodor's organisation of markers is indicated by the replacement of parentheses by a semicolon. This arrangement provides for a computational analysis of all the markers for the word "trunk". In section 2.1.4.2 it will be observed that further extensions of marker logic are appropriate for certain words.

After the emendations have been made, the question of syntagmatic relationships remains. While marker theory may be applied to immediate constituents, it is not amenable to words like stol in Lukjanow's Russian example (chapter 1, section 1.3.2.1.1) that require the scanning of the context beyond the sentence for disambiguation. In a text where 'He sat on a trunk' occurred, the establishment of a link between trunk and a word in another sentence would require an independent approach. Such an approach comes within the province of the more sophisticated classification schemes, which will be discussed in chapter 3, and which in turn come within the scope of a computerised semantics. Katz and Fodor's tree when revised would be a useful starting point for a computerised dictionary.

To organise one accurately it is necessary to distinguish

between the core and peripheral meaning of a word.[16] The former is what distinguishes it from another word. The core meaning of fork, for example, consists of semantic components of the type physical object, artifact and used for eating, and its peripheral meaning of the components - having a certain average size and not being used in Asiatic cultures. This distinction is relevant insofar as a semantic tree based on a word's core meaning is less likely to need adjusting for each new sentence to which it is applied than one based on the peripheral meaning of a word. The use of the word pen, which may mean either a writing instrument or an enclosure for animals, may be considered in the sentence 'The horse is in the pen'. It is theoretically possible to disambiguate pen by assigning to its first meaning the marker /+compact/ and to its second meaning, /-compact/, which would likewise apply to the word horse. While these markers may disambiguate pen in the above sentence, there is little guarantee that they will be equally effective in the computational analysis of unknown sentences. In the dictionary entry for pen the attributes pertaining to each of its two meanings would be indicated. Ink and cartridges, for example, would be specified as what pen (the writing instrument) contains. In the entry for pen (the enclosure for animals) animals would be specified as being contained in it. That the latter meaning of pen is the contextual one in the above sentence 'The horse is in the pen' would be determined by matching the attributes pertaining to horse with those applying to pen.

The core meaning of a word may be arrived at by examining its figurative usage, since figures of speech are often not formed

arbitrarily but on the basis of some part of a word's meaning however inapparent it may be. That obstacle is a core meaning of the word line is determined by a metaphorical analogy with the word fence. Because of this analogy with fence in 'He leapt over the fence' the use of line in the sentence 'He leapt over the line' is not anomalous.[17] By testing the word with many other metaphorical analogies one may piece together the various strands that contribute to the word's core meaning. A dictionary entry so constructed would be applicable to the translation of idioms based on figurative usage, such as those in English that involve animal names to describe personality traits. The translation of the idiom "He is a rat", for example, would be accomplished by pinpointing through the word He the contextual core meaning of rat, namely, that it refers to an unpleasant person. Thus, where the target language is one like Zuni, in which animal figures of speech are used to describe a person's physical rather than psychological characteristics,[18] the translation of rat would not be another animal name, but whatever corresponds most closely to the core meaning of the English term. Thus the figurative use of rat is incorporated into computational analysis by making it an integral part of the structure of the dictionary entry, in which the requirements of translation are met with an independent distinguisher for the psychological meaning of rat. This treatment of it corresponds to Hirschberg's[19] suggestion that "Un sens sera donc une correspondence entre une désignation dans une langue et une désignation dans une autre...."

2.1.4 The flexibility of natural language, which allows authors to

redefine words and thereby cancel in some cases the lexical relationships that would otherwise occur between them, complicates the construction of a dictionary entry. In an unsophisticated lexicon the antonym relation between freedom and slavery, for example, would be represented by the markers, /+liberty/ and /-liberty/ or the equivalent. In George Orwell's novel, Nineteen Eighty-Four, the motto 'Freedom is slavery' while cryptic is not anomalous,[20] since a penetrating analysis of the context will reveal the missing indices prominent in its explicit paraphrase 'Freedom of the body is slavery of the mind'. While this degree of accuracy in the construction of a lexical entry may not be required for mechanical translation, it would be pertinent to forms of fact retrieval that imitate the human comprehension of a text.

The utility of a dictionary depends upon how a grammar is applied to it. While the penetrating analysis of the above sentence will probably remain within the sphere of literary research, it is within the present scope of computational linguistics to provide an explicit paraphrase of the type '+freedom of thing A is -freedom of thing B', which is sufficient to convey the grammaticality if not the meaning of the above sentence. Such computation is undertaken by researchers in artificial intelligence and will be the topic of later discussion.

2.1.4.1 Weinreich's[21] claim that the relationships between the components of a sentence may also occur between the elements of an encoded dictionary definition of a single word is valid for Katz and

Fodor's markers. These are in fact adjectives belonging to order classes and are based upon a pragmatic syntax similar to that of coordinate retrieval, which will be discussed in chapter 3, section 3.2.1.2. Order classes are formed by the type of adjectives that occur in the phrase 'ten big young men'. While the English rules of grammar require the words to be in this sequence, the demands of comprehension do not. The interpretation of 'men young ten big', for example is unambiguous. On similar grounds markers of the type /+female, +offspring/ or /+offspring, +female/ for the word daughter are subject to only a single interpretation and constitute what Weinreich calls a cluster.

Groups of ordered markers which are grouped into constituent structures are called "configurations" by Weinreich[22] and "downgraded" constructions by Leech.[23] While the study of syntactic relationships is often relegated to grammar, by their appearance in the definition of a word, they also confront the lexicographer. Katz and Fodor's theory does not accommodate ordered markers. The representation of employer, for example, would be in terms of the markers /+human, +hiring/. Since this cluster would be equally applicable to the word employee for which the most appropriate paraphrase is 'a worker who is hired by someone' in contrast to 'a person who hires someone' for employer, the notation used in marker theory is overwhelmed.

In Weinreich's[24] notation the distinction between the two words is made by the direction of the arrow. The formula /human←hiring/ is the representation of the word employer and /human→

hiring/ of employee. By this convention the markers human and hiring denote the core of meaning that the two words have in common and the different arrows specify how it is organised differently in each word. For the representation of an utterance in a text the same type of notation would apply so that the phrase 'overworked employer', for example, would be assigned the formula /overworked, human←hiring/ in which the adjective-noun relation between the words overworked and employer is designated as a relation between the components overworked and human. Weinreich's notation serves to emphasise that the differentiation made by some linguists between the representation of a text and that of a dictionary definition has more to do with keeping separate the linguistic disciplines of grammar and lexicography than with linguistic reality.

In the assignment of markers to words, the results of morphological analysis, an adjunct to grammar, differ from those of lexicographical analysis, which has to do with the semantic representation of both marked and unmarked categories. The four sentences 'I counted the boys', 'I counted the boy', 'I counted the crowd' and 'The crowd is facing us' may be considered.[25] In the assignment of the marker /+plural/ to counted and boys and of /-plural/ to boy the two approaches agree. In the representation of crowd, however, they conflict. Morphological analysis designates the word as /-plural/ on the basis of the zero presence of a plural morpheme, while lexicographical analysis places it as /+plural/. Although the latter analysis would take into account the selectional restrictions according to which 'I counted the boy' is anomalous and

'I counted the boys' is grammatical, the former one embraces the sentence 'The crowd is facing us', in which crowd, being the subject rather than the object, is singular. The integration of the two types of analysis has been included within transformational grammar, in which, according to Postal[26] "the sub-component of syntactic rules which enumerates underlying phrase markers (for example, Noun Phrase and Verb phrase) is itself divided into two elements, one containing phrase structure rules (for example, sentence→Noun Phrase + Verb Phrase, Noun Phrase→Determiner + Noun) and the other containing a lexicon or dictionary of highly structured morpheme entries which are inserted into the structures enumerated by the phrase structure rules".

The weakness of marker theory, that it does not show how the semantic content of a word is organised in terms of its grammatical status, is avoided by the transformationalists. The sentences 'Pity excites the boy' and 'The boy excites pity' may be considered. In each the meaning of excites is different, being paraphrasable in the first sentence by the utterance 'stirs excitement in' and in the second, by 'causes something to be excited (in someone)'. For the first meaning of excite the marker format for the verb frighten proposed by Chomsky[27] and Postal,[28] /+Verb, +[Abstract]subject, +[Animate]object/ would be appropriate to disambiguate it from the second. In this type of formula the unbracketed categories are syntactic ones, in which the first indicates a word's part of speech, a verb in the case of excite, and subsequent categories, its syntactic environment. The categories in square brackets designate the required

semantic content of a word that is to occur in a given part of the environment. In the sentence 'Pity[Abstract] excites the boy [Animate]' the meaning of excite is deduced to be 'stirs excitement in' not just because pity and boy belong to the appropriate semantic categories, but because they assume the correct syntactic roles.

With integration of marker logic with syntax, the categories in a dictionary entry are no longer analogous to order classes. The relative positions of +[Abstract], subject, +[Animate] and Object in the above formula are important, since a different sequence such as the one in the formula /+Verb, +[Animate] subject, +[Abstract] Object/ would designate the second meaning of excite, namely, 'causes something to be excited (in someone)'. This type of notation, which specifies the meaning of a word in terms of its environment, is the crux of a computerised semantics and will be discussed further in chapter 4.

2.1.4.2 While a variety of syntagmatic relationships other than the one between the subject, verb and object might be included in a dictionary entry, it is not within the scope of the present development of linguistic theory to provide an exhaustive list of them. A sample, however, will suffice to emphasise the necessity of a more elaborate assemblage of indices than that provided by Katz and Fodor's marker theory.

One syntagmatic relationship concerns verbs of mention, of which speak in the sentence 'It is nonsense to speak of a king as made of plastic' is an example.[29] Within the framework of Katz and Fodor's

theory the words king and plastic would be abstracted from the sentence and through their respective formulae, "king"= +animate [monarch] or -animate [chess piece] and "plastic" = -animate; the marker logic outlined in section 2.1.2 would ensure that king meaning a chess piece would be selected as the contextual meaning. Since the usually anomalous meaning is the one required in the environment of 'It is nonsense to.....' marker logic would work only if in this environment the formula for king were altered by a grammatical rule to "king"= -animate[monarch] or +animate[chess piece]. Such an alteration might take place in mechanical procedure through the assignment to the word nonsense of a symbol, which would be operative whenever the word speak introduced a noun phrase syntactically connected with nonsense.

The application of mechanical procedure may be complicated by the absence of a verb of speaking. Such an omission is evident in the sentence 'That stallion is a mare', which - as a facetious remark - is not anomalous.[30] The missing indices may be observed in the paraphrase, 'What you called a stallion is a mare'. In the present state of computational linguistics, the mechanical detection of such facetious remarks will have to be shelved, although a pragmatic measure may be adopted. Since in written works apparent contradictions usually occur on purpose, they might in default of any other analysis be treated as cases where a verb of mention is implied, where they occur frequently.

The scope of the amended marker notation may be extended to

cases where three syntagmatic links are involved. The relationship represented in the above formulae by the operator or is appropriate in the representation of the anomalous use of the word sad in the sentence 'John is as sad as the book he read yesterday',[31] for which the formula for sad is "sad" = /-animate[cause emotion] or +animate [have emotion]/. The or is the same exclusive one that featured in the one-dimensional notation for canard in Katz and Fodor's example (section 2.1.2). In the above sentence the marker /+animate/, is identified as representing the contextual meaning of the word sad, because the marker for John, /+animate/ which Weinreich[32] calls a transfer feature, is matched against the formula before the one for book, /-animate/.

The or operator is not applicable to all cases of three syntagmatic links. In the sentence 'John is heavier than this rock',[33] the grammaticality of the use of 'heavier' might, for example, be conveyed through the formula, "heavier = /+animate, -animate/, where the comma functions as the antonym of or. Both markers may be selected as appropriate for a given context, since whatever type of noun is qualified by this word, heavier has the same meaning. A borderline case is provided by the word take. While generally its representation by two markers separated by the operator or is accurate, this type of representation does not take into account the occurrence of zeugma, whereby the use of take in the satirical sentence 'Queen Anne does sometimes counsel take and sometimes tea', for example, is permissible. A possible formula for the word take that takes into account this usage might be "take" = /+zeugma/ or

//, /-zeugma/, where the double slash lines would indicate that the choice of operator depended on the extralinguistic context of take.

The constraints of syntactic usage affect the plus and minus indicators. The dichotomy between them is not suitable for the syntagmatic representation of all the relations that involve mutually exclusive markers. These come under Lyons'[34] heading of incompatibility. According to his definition "the assertion of a sentence containing one of the terms over which the relation holds can be shown to be understood as implicitly denying each of the sentences formed by the substitution of any one of the other terms of the set in the context in which the given term occurs." Katz and Fodor's indicators are suitable for expressing the relationship between antonyms (polar systems) but not that between words in a multiple taxonomic grouping like the colour system.[35] That the indicators do not effectively represent it may be observed in their lack of ability (however little required in practice) to detect such sentences as 'Red is green' and 'Blue is green' as contradictory.

Multiple taxonomic systems may be divided into two kinds, hierarchic and non-hierarchic.[36] An example of a non-hierarchic system may be observed in the names of colours, for which instead of two indicators there would be several. For the words red, green and blue, for example, the markers might consist of the respective combinations of indicator and descriptive, /Red Colour/, /Green Colour/ and /Blue Colour/. The anomalousness of the sentences in the previous paragraph would be detected by the same marker logic as

before, since Red, Green and Blue oppose each other as plus and minus do, but through a larger vocabulary of indicators.

An example of a hierarchic system may be observed in the relationship between part and whole in body parts, among which the words man, arm and finger[37] may be considered. This system demands more changes in the marker code than the non-hierarchic one above, since not only are more indicators necessary but they need to be hierarchically ordered. To express the hierarchy between the words man, arm and finger markers of the type /3 Body/, /2 Body/ and /1 Body/ might be assigned respectively to them. The sentences 'The man has an arm' and 'The arm has a finger' would be recognised as being more acceptable than 'The arm has a man' by the fact that the subject of the sentence is assigned a higher number than the object in the first two.

2.2 The Semantic Interpretation of Syntactic Structure

2.2.1 While the nucleus for research on the construction of dictionary entries was provided by Katz and Fodor's[38] marker theory, no theory of equal weight has appeared as a mechanical model of how syntactic structure conveys information. A minimal requirement of such a model would be to illustrate how sentences consisting of different elements may be synonymous. Some of the research based on Chomsky's[39] transformational grammar, while an attempt to meet this requirement, focuses only on the least complicated problems of paraphrase. Categories used as a tool by which to apply rules for

recognising paraphrases are called cases or role indicators.

One kind of paraphrase involves the replacement of one word by a certain other in conjunction with a reversal of word order. The relationship between these two words is said to be a conversive one.[40] By such criteria the sentence 'John sells books to Mary' would be recognised as a paraphrase of 'Mary buys books from John' because of the interchanging of the words Mary and John and the conversive relationship between sells...to and buys...from. A convenient notation might be /Active barter/ for sells to and /Passive barter/ for buys...from, in which the non-capitalised item is a marker and the capitalised one is a case. A transformational rule would recognise the above two sentences as paraphrases by means of the conversives, which match with respect to markers but differ with respect to cases, one being Passive and the other, Active. This type of notation is applicable to Chomsky's familiar active-passive transformation, so that a sentence like 'Joe strikes John' may be recognised as the paraphrase of one like 'John is struck by Joe' through the assignment of /Active hit/ to strikes and /Passive hit/ to 'is struck by'. For the sentences in this example the setting up of the notation is aided by explicit indices.

Cases[41] are applicable to the types of paraphrase that do not involve a paradigmatic relationship like the conversive one above. The fact that the sentence 'John ruined a table' may be paraphrased by 'What John did to a Table was ruin it', but not the sentence 'John built a table' by "What John did to a table was build it' may be

traced to the different semantic categories to which ruin and build belong. In order that a verb may fit into the slot after 'What John did to the table was.....', the existence of the object of the verb must be predicated as being prior to the action of the verb. The verb ruin but not build meets this requirement. Correspondingly the category of ruin is specified as Affecting and that of build as Effecting, in which capitalisation designates these representations as cases.

In Nida's[42] "object - event" analysis, cases are used to correlate words of different parts of speech and on different syntactic levels to detect paraphrases. One may consider phrases consisting of an adjective and noun that are synonymous with combinations of verb and adverb, in which the adjective is isomorphic with the adverb and the noun with the verb. The three words in Nida's sentence 'He works excellently' and the three respective non-bracketed words in 'His work [is] excellent' that correspond to them are amenable to case analysis. Three cases in the formula, Object⟶ Event ⟵ Abstract, in this order represent the words of each of the above sentences. For computation, these cases would have to be related to a surface structure grammar, which Nida does not do. The atomisation of parts of speech, in this case adjective, noun, verb and adverb, which Bar-Hillel[43] aimed for in his categorial grammar, might provide the requisite components for arriving at the formula through step by step procedure.

Fillmore's[44] case system covers paraphrases that have to do

with transitive and intransitive verbs, of which the word move is an example of both. A syntactic analysis of the sentences in his example, 'The rock moved', 'The wind moved the rock' and 'I moved the rock with a stick', would indicate rock as the subject in the first sentence and the object in the second and third. Since the combinations of the words moved and rock bear the same meaning throughout, notwithstanding word order, the word rock in Fillmore's notation is assigned the case, Object. The words stick and wind, since they refer to a force directly responsible for action, are given the case, Instrument. The word I, which refers to a force indirectly responsible for action, is assigned the case, Agent.

The interaction between the cases is indicated through Fillmore's following formula: — Object; (Instrument); (Agent), in which the brackets signify that the appearance in a sentence of the case in question is optional. The rule according to which words belonging to a given case fit into a sentence is as follows: if the Agent is not present, the Instrument is the subject and if this case is not present the Object is the subject of the sentence. Even with the indication of the surface structure to which the cases relate, Fillmore's system is incomplete. In order to apply the case system the formula for each word must include markers. For example, the noun rock, since it cannot function as the Object of every verb, might be assigned the formula /Object; motion/, and the verb move might be assigned the semantic component /motion/. The fact that rock and move have the same semantic component, /motion/, and that rock is assigned the case, Object is the computational means by which rock is

detected as the Object of move.

2.2.2 The resources of theoretical linguistics from which a computerised semantics may be constructed peter out beyond the recognition of the simplest cases of paraphrase, although further lines of investigation have been postulated. Bellert,[45] for instance, claims that "The surface structure of an utterance in which linguistic indices are explicitly expressed is clearly much closer to its LS [Logico-Semantic] structure representation than those of its paraphrases in which linguistic indices do not occur, although they are somehow implied if the utterances are recognised as paraphrases". The LS structure of an utterance is the formulation of its meaning in terms of explicit indices. The syntagmatic difference between the two sentences 'John is easy to please' and 'John is eager to please' would be conveyed in LS structure through the paraphrases 'John is easy for someone to please' and 'John is eager to please someone', which contain the indices for and someone. In computational linguistics these techniques of transformational grammar have been applied in the construction of an interlingua.

Another line of investigation is pursued by Weinreich[46] in his formulation of certain syntagmatic properties of utterances. These properties come under his heading of linking, within the scope of which are included the sentences 'The wall is white', 'The wall's whiteness is astonishing' and 'The wall is astonishingly white'. They are represented by his formulae, (wall, white), (wall, white astonishing) and (wall, astonishing white), in which the adjectives

represented in superscript are considered to form the most direct link with wall and the others, a less direct one. Weinreich's notation conveys information that would be provided by immediate constituent analysis, in which the second sentence would be said to differ from the third in having wall's and whiteness rather than astonishing and white as immediate constituents. The originality of Weinreich's contribution appears to consist mainly in his placing syntagmatic linking within the province of semantics.

In the analysis of how information is conveyed via syntactic structure linguists have focussed their attention upon the task of representing the constituents of sentences themselves and have tended to shy away from incorporating encyclopedic knowledge. In Nida's[47] terminology, the first task concerns constructions from which the meaning of a whole utterance can be derived from the meaning of its parts, semantically endocentric ones, and the second task, constructions from which the meaning of the whole utterance cannot be so derived, semantically exocentric ones. Nöel's[48] utterance 'libération du prolétariat', which he considers to be an example of "metaphorical logic", is an exocentric construction, since from the dictionary meaning of the words libération, du and prolétariat it is not possible to derive the paraphrase 'établissement d'un nouvel ordre économique favorable aux travailleurs'.

The reliance of authors on extralinguistic context makes paraphrase recognition in the case of exocentric constructions difficult. Because of such reliance the dictionary meaning of a word

in a given context may be rendered redundant. For example, while the two sentences 'A paradigm is a set of substitutable forms' and 'By a paradigm we understand a set of substitutable forms'[49] differ superficially with respect to the lexical meaning of understand, they are nonetheless paraphrases of each other, because this word conveys no information in its context and effectively constitutes a function word. In fact retrieval it would be desirable to have such a word rejected as a keyword for this reason.

The question of encyclopedic knowledge is raised by Todorov[50] in his discussion of how to categorise the different connotations of the French word manger in the utterances 'manger la soupe' and 'manger une pomme'. In the first, manger refers to the act of eating by means of a spoon and in the second, by means of one's hands. Todorov calls these differences ones of reference rather than ones of meaning. In mechanical translation the common core of meaning that the two uses of manger have in these utterances, namely the act of putting something in one's mouth and digesting it, would justify treating the word as being unambiguous, if the target language were a language like English, in which the word eat has similar connotations. In sophisticated form of fact retrieval involving the treatment of relationships beyond the sentence level by means of a classification scheme the connotations of manger would probably be differentiated.

What Laffal[51] calls experiential validity comes under the heading of encyclopedic knowledge. The two sentences 'I talked with you' and 'I will talk with you' may be compared with 'I agreed with

you' and 'I will agree with you'. Between the first two sentences there is a temporal distinction and the word will refers to the future. In the second group of sentences the future tense usually expressed by will does not have experiential validity, for, while one may set a date as to when one will talk in human culture, one does not usually prophesy one's agreement with someone.

The representation of encyclopedic knowledge calls for greater specification than the type of cases mentioned so far provide. Fillmore's[52] cases indicate the semantic relationships between words only insofar as these relationships show how a sentence of one kind of syntactic structure may be paraphrased by a sentence of another kind. Schank[53] appropriately pointed out that of Fillmore's sentences (section 2.2.1) 'The wind moved the rock' and 'I moved the rock with a stick', the first differed from the second in that the act of blowing was implied, which Fillmore's case system for I and wind did not indicate. For Katz and Fodor's[54] sentences 'Should we take junior back to the zoo?', 'Should we take the lion back to the zoo?' and 'Should we take the bus back to the zoo?', semantic components such as Animate and Human for the object of each sentence would serve to indicate that the word take has different implications in each sentence, but not how. In order to register certain information, namely that lions are kept in cages, buses are ridden on and that humans visit zoos, the dictionary entry for take back would have to incorporate encyclopedic knowledge.

2.2.3 The notations of Quillian's[55] mechanical memory and

Schank's[56] mechanical intelligence are able to accommodate both linguistic and encyclopedic knowledge, since they represent a different type of relationship from that which was the predominant concern of theoretical research. This difference may be brought out by the explanation of certain terminology. In linguistics, relationships between words in a text are said to be "syntagmatic" or "synthetic" and those between dictionary entries, "paradigmatic" or "analytic". Examples of the latter are the relationship between fixes and teeth, which constitute part of the definition 'Someone who fixes teeth' of the word dentist, and the one between rich and poor, as they occur as dictionary entries. Synthetic relationships are those that occur between words in a text and which do not form part of any dictionary entry. For the purpose of explication it would be convenient to amend the above terminology so that the term "analytic" refers to words that are part of a dictionary definition and the term "paradigmatic" to relationships between dictionary entries.

2.2.3.1 The empirical approaches of Schank and Quillian have to do with analytic relationships, whereas theoretical linguistic research is concerned with paradigmatic relationships. Quillian's[57] memory consists of nodes connected by different kinds of links. Each node represents one of the meanings of a lexical word and each link designates a function word, which conveys syntactic relationships between lexical words. Dictionary entries are represented by "type nodes" and the lexical words that are part of their definition, by "token nodes". The word plant, for example, would be assigned three type nodes, PLANT 1, PLANT 2 and PLANT 3 and their respective meanings

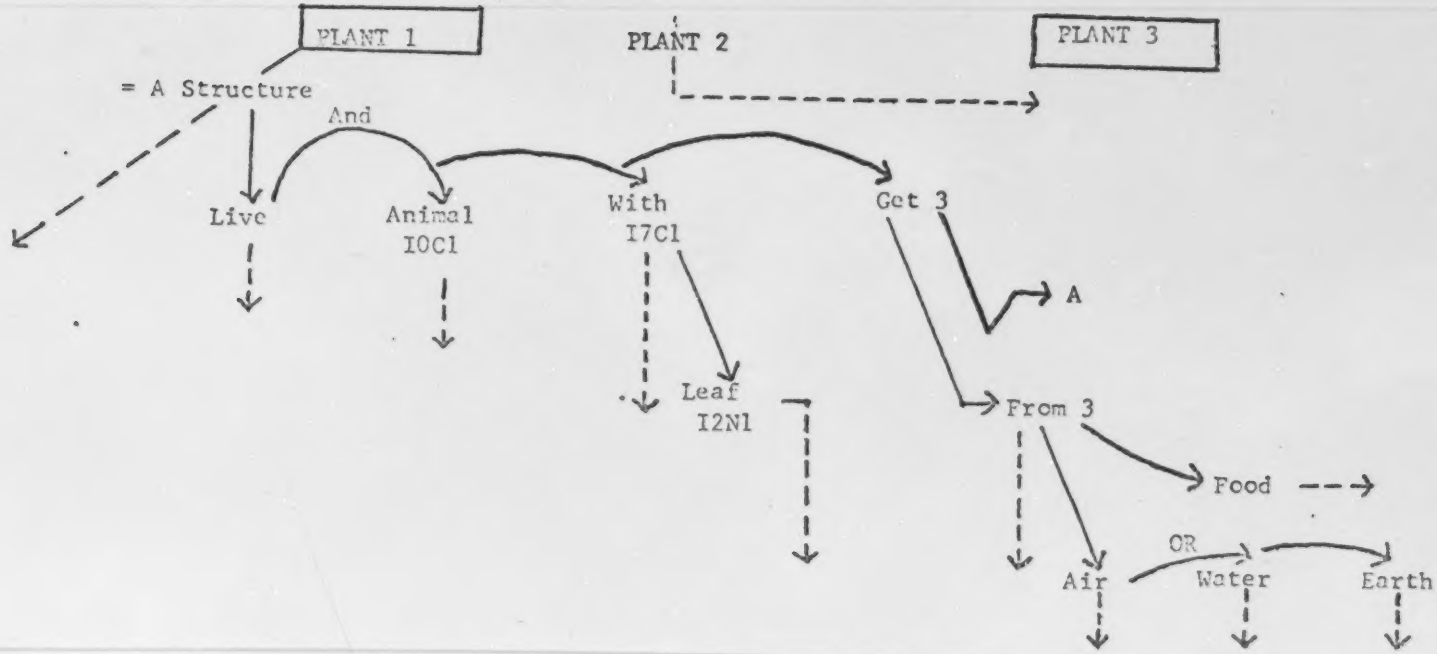


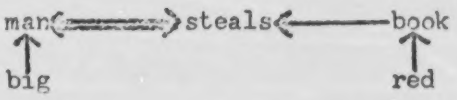
Figure 2: Quillian's memory.

'botanical organism', 'fix something firmly' and 'industrial complex', would be encoded into token nodes and links of the type provided in figure 2. In this model the full meaning of a word is derived from an exhaustive tracing of the definitions of its token nodes. A word so traced is called a patriarch word by Quillian. In the above diagram Plant 1 would be a patriarch word if its full meaning were searched for by tracing the dotted arrows which lead from each of the token nodes.

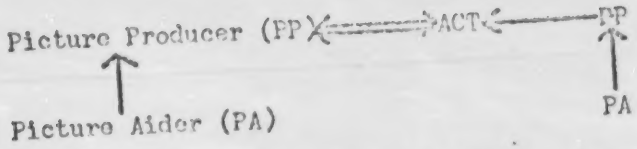
While the mechanical memory has mainly to do with analytic relations, paradigmatic ones could be represented implicitly by a careful formulation of the definitions for each word. Ceccato's[58] model of semantic relationships represents paradigmatic relationships, such as those of hyponymy, antonymy and conversiveness by a diagram consisting of words linked by correspondingly numbered arrows. In his notation the hyponymy relationship between Animal and Dog would be conveyed in the form 'Dog $\xrightarrow{03}$ Animal', in which the species occurs before the arrow and the genus after it and where 03 designates the hyponymy relationship. In such a diagram the two utterances pear tree and tree would not be included, since the hyponymy relationship is conveyed by explicit indices and thereby is within the province of grammar. Nonetheless the same relationship holds between these utterances that occurs between Dog and Animal. Tree is the genus and pear tree, the species. In the memory it would not be difficult to incorporate the hyponymy relationship if dog were assigned explicit indices in the form of the phrase canine animal, which would then be encoded. The analogy between the first two utterances, pear tree and

dog, and the second two, tree and animal would not be obscured through their separation into two linguistic disciplines, which would happen in Ceccato's model, unless it incorporated an indefinitely large number of phrases and possibly of sentences.

2.2.3.2 In Schank's[59] mechanical intelligence the notation consists of terms which, for convenience, he represents with English words rather than with code numbers, and of relationship items portrayed in the form of different kinds of arrows. A sentence such as 'The big man steals the red book' would receive the following two-dimensional representation:



The double arrow represents what is called a two-way dependency relationship and the single arrows, one-way dependency relationships. This representation parallels immediate constituent analysis insofar as the upward single arrows designate linked words, the horizontal single arrow denotes linked elements of which one is a phrase and the double arrow designates the linking of phrases, in this case, 'big man' and 'steals the red book'. This type of notation, which appears again in Wilk's approach (chapter 4), is more than an I.C. analysis in that his components are words that have undergone language normalisation. These words are parsed according to their deep structure significance so that the representation of the above diagram, in which each term is replaced by its concept class, is as follows:



The connection between a sentence and this type of conceptual representation is provided by "English realisation rules" of the following type:

PP⟹ACT = Noun Verb, PP⟵ACT = $\left\{\begin{matrix}\text{the}\\ \text{a}\end{matrix}\right\}$ Adjective / - Noun.

The above representation is thereby observed to correspond with the parse, $\left\{\begin{matrix}\text{the}\\ \text{a}\end{matrix}\right\}$ Adjective / Noun Verb $\left\{\begin{matrix}\text{the}\\ \text{a}\end{matrix}\right\}$ Adjective Noun, of the sentence 'The big man steals the red book'.

In the dictionary the above types of normalised words are grouped into semantic categories. The representation for one meaning of the word ball, for example, would be as follows:

PA	PP	ACT
has texture	located anywhere	rolls
has colour	for anyone	bounces
has beauty	belongs to people	hits
.	.	.
.	.	.
.	.	.

An entry so structured is equipped for the resolution of multimeaning. The concept classes, PA, PP, and ACT represent the semantic attributes with which adjectives, nouns and verbs respectively in the environment of ball must be compatible, if this meaning of ball is to be selected for a given context. Examples of words that meet these criteria are to be found in the sentences 'The ball is red', 'The ball is John's' and 'The ball is rolling', in which ball is understood to refer to a round object rather than to dancing. In order to effectively judge the above type of file, the results of large scale applications will

have to be examined. However, the categories in it appear to be based on intuition and it may be suspected that the notation for them may resemble that of Booth's concept numbers.

Like the nodes and links in Quillian's[60] memory, the concepts and relationship items convey paradigmatic relationships implicitly through analytic ones. The relationship between cat and edible would be conveyed through the representation of edible as follows: one⇐=beat⇐='thing' (N). Whereas in Quillian's memory the semantic content shared by a group of words was elicited by tracing the nodes and links, in Schank's[61] dictionary the meanings are factored into particular files. For example, although some of the concepts appropriate to the word ball would be found in its entry, others would be found in a "physical object" file. Files would list encyclopedic as well as linguistic information. For example, to understand the sentence 'Did Nixon run for President in 1964?' a machine would search the experience files and interpret President as 'President of the U.S.'

While Schank's mechanical intelligence accommodated encyclopedic knowledge, research in computational linguistics was generally orientated towards the traditional dictionary rather than the encyclopedia. The makers of classification schemes, who were often concerned with specific subject areas, had a different perspective. For example, from a chemist's point of view a suitable classification of a document entitled 'The Conversion of Water into Hydrogen and Oxygen by Hydrolysis' would probably involve for the

utterances water, hydrogen and oxygen and hydrolysis the categories, starting material, product and process respectively, which would contribute to the structuring of encyclopedic knowledge. For source material on how to organise it, it would be appropriate to examine the classification schemes.

Footnotes to Chapter 2

1. Katz and Fodor 1966
2. Schank 1969
3. Katz and Fodor 1966
4. Bally and Sechehaye 1959
5. Prieto 1964
6. Prieto 1964
7. Katz and Fodor 1966
8. Lyons 1963
9. Weinreich 1966
10. Duchacek 1965
11. Katz and Fodor 1966
12. Larousse dictionary 1960
13. Postal 1964
14. Nida 1963
15. Bolinger 1965
16. Bierwisch 1969
17. Fillmore 1969
18. Nida 1963
19. Hirschberg 1964
20. Leech 1969
21. Weinreich 1966
22. Weinreich 1966
23. Leech 1969
24. Weinreich 1966
25. Weinreich 1966

26. Postal 1964
27. Chomsky 1965
28. Postal 1964
29. McCawley 1968
30. Leech 1969
31. McCawley 1968
32. Weinreich 1966
33. Chomsky 1965
34. Lyons 1963
35. Lyons 1963
36. Leech 1969
37. Chomsky 1965
38. Katz and Fodor 1966
39. Chomsky 1965
40. Chomsky 1965
41. Fillmore 1968
42. Nida 1963
43. Bar-Hillel 1953
44. Fillmore 1969
45. Bellert 1969
46. Weinreich 1966
47. Nida 1963
48. Nöel 1968
49. Bierwisch and Kiefer 1969 .
50. Todorov 1966
51. Laffal 1970

52. Fillmore 1969

53. Schank 1969

54. Katz and Fodor 1963

55. Quillian 1967

56. Schank 1969

57. Quillian 1967

58. Ceccato 1961

59. Schank 1969

60. Quillian 1967

61. Schank 1969

3 CONTRIBUTIONS OF THE CLASSIFICATION SCHEMES

3.1 General Classifications

3.1.1 As was demonstrated in the previous chapter, a classification of encyclopedic knowledge and the means of showing how to encode utterances of natural language into it would be essential to a computerised semantics. Various classifications exist and are of two kinds, general and special. The special classifications, each of which embraces only a small part of the spectrum of knowledge but in great detail, were created to serve the particular needs of researchers in a given field and were significant in their abandonment of the principles on which the general schemes were based. The general classifications, which embrace the whole spectrum of knowledge, were created for use in library science, in which the encoding of documents to locate them on the shelves, is performed by human intuition. The notation of the general classifications is correspondingly not useful for the mechanical encoding of utterances. Nonetheless, all these classifications provide source material and even the notation offers lessons, albeit negative, for computational linguistics.

In a classification scheme there are usually three basic components: schedules, which are lists of groups of symbols that can be suffixed to the main notation, the general tables, which provide encyclopedic data on the lexical words of a language in the form of a code-English dictionary and thereby material from which to construct idioglossaries, and an index, which is an English-code dictionary.

Because the scope of a classification scheme includes encyclopedic data, the assessment of it will be different from that of the usual dictionary. Brown[1] in his Subject Classification who aimed to classify the words in his index as concrete unities missed this point. For the word eggs, for example, he provided only one code number, F601. The postulated unity, however, lies solely in the fact that the word refers to the hard-shelled reproductive body of a fowl or bird. Eggs fits into many subject areas. In Dewey's[2] Decimal Classification each application of the word is represented by a code number. For the context of nutrition, for example, eggs is represented by the number 612.39283 and for that of ornithology by 598.2.

Among general classifications the following will be treated; the Library of Congress (LC), the Dewey Decimal (DC), the Universal Decimal (UDC), the Bibliographic (BC), the Subject (SC) and the Colon (CC) classifications. UDC[3] and DC[4], the first of which is derived from the second, may be grouped together. In both, knowledge is fitted into the following decimal framework; 000 generalia, 100 philosophy, 200 religion, 300 social sciences, 400 language, 500 science, 600 technology, 700 fine arts, 800 literature and 900 history, travel, biography. Successive subdivisions of these topics are made by inserting digits between one and nine in the tens and units columns and thereafter beyond the decimal point. For example, technology is 600, engineering is 620, mechanical engineering is 621 and machine tools is 621.9. While the decimal point is not essential, it is inserted to divide up the digits for the human eye.

The zeros at the end of a number, which are not necessary, are eliminated in UDC.

While the above notation is compact, the objective mapping of knowledge seems to have been sacrificed to accommodate it. The reason for grouping medicine 610, agriculture 630 and building 690 under the catch-all term, technology 600, appears to be that there are not enough digits in the decimal system to accommodate more main classes. Rangathan's[5] octave device provided a means of overcoming this limitation. In his notation a digit preceded by any number of nines is considered to belong to the same subdivision as one preceded by no nines at all. The numbers, 1, 2, 8, 91, 92, 93, 98, 991 and 992, would represent main classes. As the number of nines that can precede a digit is infinite, so is the number of possible terms in a given subdivision.

While DC and UDC share a common notation they differ in the means of building flexibility into it. In DC flexibility is provided by "divide-like" instructions, through which the same string of digits may be segmented differently to convey different information, cross-references being made from one part of the general tables to another. An example of the use of the instructions may be seen in the classification of Proverbs 398.9 according to the language in which they are written. While digits may be added in the way described previously to this number to denote the subdivisions of Proverbs, such an approach would fail to utilise the ready-made categorisation of languages set up for another subject, that of linguistics. The subdivisions of language are represented by the

numbers 420 to 490. Therefore, to subdivide Proverbs the instruction 'divide like 420 to 490' is provided, which means that the number for a particular language is attached to the one for Proverbs but without the initial 4. The number for Indo-Iranian Proverbs is 398.9911, which is a conflation of Proverbs 398.9 and Indo-Iranian language 491.1. While in the general tables the "divide-like" instruction is applied to increase the inventory of terms with the same notation, the device could be extended indefinitely to represent synthetic relations as well.

The capacity to represent synthetic relations is provided in UDC by auxiliary symbols. The most prominent of these is the colon, which has the same meaning that the word and has in English. This symbol allows strings of digits to be regrouped without a change in their meaning. For example[6], 66 chemical technology and 658 industrial management can be synthesised into either 66:658 or 658:66 to denote 'management in the chemical industry'. The infixing of one string inside another by means of square brackets provides for further flexibility of representation. A grouping of given documents by numerical order might occur as follows: 620.191: 669.3 'discoloration of copper', 620.191: 669.4 'discoloration of lead', 620.192: 669.3 'swelling of copper', or as follows: 620.19[669.3]1 'discoloration of copper,' 620.19[669.3]2 'swelling of copper', 620.19[669.4]1 'discoloration of lead'. The first grouping collects documents on discoloration and the second, those on the defects of copper. Because of its provision for alternative ways of grouping documents and of its capacity to convey information about them in very

great detail, UDC is not just a library classification but also an information retrieval system.

Since UDC is tied to the enumerative framework of DC, the provision for alternative grouping is not consistent. This fact is underlined by the apparent redundancy of certain digits in the notation. In 633.15 - 272.6; 632.937 'Biological Control of Injuries caused by Locust Pest to Corn', which is synthesised from 633.15 - 272.6 'Injuries caused by Locust Pest to Corn' and 632.937 'Crop Protection by Biological Control',[7] the digits 6 and 3 on both sides of the colon signify twice over that the document comes under the category of agriculture. This apparent redundancy is unavoidable, since these digits are needed as null terms, as the number 2.937 by itself has a different meaning from 2.937 in the string 632.937.

Bibliographic Classification (BC) and Subject Classification (SC) provide means for representing synthetic relations. In BC the comma serves the same function as the colon in UDC. The title 'Protection of Corn against Locust Pests in India in 1967', for example, would be represented by the components UA Agriculture, QT Corn, JQDL Locust Pest, H Protection, q India and U Recent Period in the formula UAQT, JQDL, Hq, U.[8] In SC, the boundary between one string and the next is marked by a change from a numerical to an alphabetic base. A title like 'Unemployment in the Shipbuilding Industry' would be represented by either B650L118 or L118B650.[9] Letters represent main classes. K, for example, signifies philosophy and religion, and groups of digits (000 to 999) represent the sub-divisions of these classes. For instance, K951 signifies Catholic

Apostolic Church and K952, Christian Endeavour Society. To intercalate new subjects, more digits are added.

The most inflexible of the general classifications with respect to notation is LC,[10] which is based upon an actual collection of documents rather than on a theoretical map of knowledge. Its main classes are as follows: A - General Works; Poligraphy; B - Philosophy, Religion; C - History, Auxiliary Sciences; D - History and Topography; E and F - America; G - Geography, Anthropology, Sport; H - Social Sciences; J - Political Science; K - Law; L - Education; M - Music; N - Fine Arts; P - Language and Literature; Q - Science, General; R - Medicine; S - Agriculture, Plant and Animal Industry; T - Technology; U - Military Science; V - Naval Science; Z - Bibliography and Library Science. Subdivisions are made by means of a second letter and four digits. While gaps between numbers provide for expansion, the notation cannot accommodate synthetic relations. Because of its lack of cross-referencing, LC may be considered as a coordinated series of special classifications, each main class being independent of the others. For example, in Statistics, periodicals and congresses are represented as HA1 and HA 9 to 11 respectively, while in Economic Theory they are respectively HB 1 to 9 and HB 21 to 29.

In the other classifications the inflexibility of notation is relieved by schedules. In SC,[11] numbers belonging to a schedule are distinguished by a preceding point. Thus I229.10 'History of Landscape Gardening' may be recognised as a concatenation of I229

'Landscape Gardening' and .10 'History for general use in all classes'. In DC,[12] a preceding zero distinguishes a number that belongs to a schedule. For example, 614.05 'periodical on public health' is interpreted as 614. 'public health' and 05 'periodical', while 505 'periodical on science' is segmented into 5(00) science and 05. In BC, where the main notational base is alphabetic, the distinction is made by means of numbers so that a concatenation of BOV and 3 History, for example, provides BOV3 'History of the BBC'. In UDC,[13] punctuation is used. For example, 624=30a is 'civil engineering written in German' and 624(47) is 'civil engineering in the U.S.S.R.'.

3.1.2 The enumerative framework of the general classifications provides for a very economical representation of information. In DC, the meaning of a digit is determined not only by the column in which it occurs, but also by what digit it follows. For example, 3 indicates a subdivision of technology after 6 in 63 agriculture but one of Social Sciences after 3 in 33 economics. Such economy is attained at the expense of the accommodation of synthetic relations, which, as was observed in the last section, are only indirectly represented. From the point of view of computation this drawback is inconvenient.

The notations of the general classification schemes were created on the principle that any one term was the species of only one genus as depicted by the tree of knowledge analogy. Such a view does not take into account the complexity of lexical organisation.

Terms are like the elements of the mathematical equation, ab + bc + ad, which may be represented as a(b + d) + bc or b(a + c) + ad, in that they can be grouped according to more than one category. For example, lamb may be placed with sheep under the category ovine or with puppy and pony under the category young.

The consequences of rejecting the tree of knowledge analogy may be observed in an adaptation of Sharp's[14] example, in which the categories into which terms are placed are explicitly indicated. In a classification of military science a term like military aeroplanes 600 might be subdivided into three other terms fighters 610, bombers 620 and Transport planes 630. Each of these might in turn be subdivided. Thus fighters 610 would embrace the terms 'single-engined fighters' 611 and similarly bombers 620, the terms 'single-engined bombers' 621, 'twin-engined bombers' 622 and 'three-engined bombers' 623. 'Transport planes' 630 would be subdivided into 'twin-engined transport planes' 632 and 'three-engined transport planes' 633. In this notation semantic categories are specifically designated by columns. The tens column represents the genus of purpose, and the units column, that of the number of engines. When a particular category is not applicable, the others may be preserved by inserting zero between other digits as a null term. Thus 'three-engined military aeroplanes' would be 603.

In the above notation the maximum capacity of representation of a three digit number is twenty-seven terms (9+9+9), whereas in DC the capacity is seven hundred and twenty-nine (9x9x9). If the

octave device is used, the difference increases. However, Sharp's notation is more attractive from the point of view of mechanical retrieval, since machine may more readily retrieve a term by identifying a digit in a specified column than one as it occurs after certain other digits.

Drawing the principle of flexibility to its logical conclusion, the makers of the special classifications create notations of unordered digits - or groups of digits (in which the octave device is applied). The terms and their corresponding symbols for the classification of military aeroplanes would be of the type, single-engined 1, twin-engined 2, three-engined 3, fighting 4, bombing 5, transporting 6, and military aeroplanes 7 and a document entitled, for example, 'Three-engined Fighters' or 'Three-engined Fighting Military Aeroplanes' would be represented by the number 347 or 743 or any combination of these digits. This type of notation is adopted by Perry, Kent and Berry in their generic encoding, in which groups of letters replace the above type of digits. The letter combinations from left to right represent the classes and subclasses into which a term fits. For example, animal is NA, mammal is NA MA, dog is NA MA DO and terrier is NA MA DO TE. This systematic sequence of letter combinations makes for human convenience and speed of computation, although no ambiguity would result from any other sequence. The pros and cons of such a notation will be discussed further in section 2 under the heading of Coordinate Retrieval.

The exception among general classifications is Rangathan's[16]

Colon Classification in that it is not enumerative. Rangathan designated his categories (or "facets" as he calls them) of which there are five, explicitly as follows: Personality [P], Matter [M], Energy [E], Space [S], Time [T], which appear in this order in the representation of document titles. Matter denotes materials, Energy indicates an operation, process or problem to be solved, Space and Time denote geographical and chronological subdivisions respectively and Personality seems to include miscellaneous information. In the representation of a title Personality is followed by a comma, Matter by a semicolon, Energy by a colon, Space by a full-stop and Time by an apostrophe. Encyclopedic information about the words that may appear in titles is incorporated into the syntagmatic framework described above by means of an index, in which the five categories are distinguished by square brackets. For example, in the entry 'lending 2 [E],62 X [P],62 [E] 1', "lending" is interpreted as belonging to class 2 library science, where it is denoted by the number 62 in the Energy facet, and to class X Economics, where it is denoted by 62 in the Personality facet and by 1 in Energy.

Various procedures govern the encoding of a title. In the first, it is parsed with the aid of an index. A document entitled 'Spraying Instrument and Chemicals to Mitigate the Virulence of the Injury to the Stem of the Rice Plant during the 1967 Dry Period in the Cauvery Delta in Madras' would be analysed into components as follows: Agriculture. Rice Plant [1P1]. Stem [1P2]. Injury [1M1]. Virulence [1M2]. Mitigation [1E]. Chemical [2M1]. Application [2E]. Spraying instrument [3M1]. Madras [S1]. Cauvery Delta [S2]. 1967

[T1]. Dry Period [T2]. In this notation the number before the letter denotes which round a word of a particular category belongs to and the number after the letter denotes its level. A round of facets is a clause consisting of Personality, Matter and Energy, any of which may be absent in any particular title. Where two or more facets of the same type, two Personalities for example, occur in the same round, they are said to belong to different levels and are correspondingly assigned different numbers. After all the rounds have been represented, the Space and Time facets are inserted. When a title has thus been parsed, numbers replace words to provide 'a title in focal numbers', which in this example is as follows: J.318 [1P1].4[1P2].4[1M1] 0c7[1M2].5[1E].3[2M1].7[2E].5[3M1]4411 [S1] e50c[S2].N67[T1] e1[T2]. In the final class number the punctuation specified earlier replaces the bracketed tags.

In the classifications of the various fields of knowledge the relationships between them are represented by relationship items (or phases) in what Rangathan[17] calls a phase analysis. This analysis is related to Micklesen's[18] division of knowledge into idioglossaries (described in chapter 1, section 1.3.2.1.3) based on the assumption that some fields of knowledge may be considered as compounds of main classes, geophysics, for example, being the 'influence of geography on physical science'. In the notation of CC the subject is represented by the formula, COgU, where C and U represent the respective main classes, physical science and geography, and where zero indicates a change of phase and the lower-case character g represents the phase of influence. Other relationship items are b

and k bias, c and m comparison and d and n difference. Rangathan's notation in providing a syntax represents a significant departure from the norm of the other general classifications.

3.2 Special Classifications

3.2.1 In all the special classifications the categories of the type described in section 3.1.2 are adopted, through which distinct dictionary entries are formed, each with its own notation so that the representation of a document is more than an idioglossary summary. These classifications may be divided into two main types. In one, the categories are represented by descriptors (or semantic factors as they are sometimes called). In this type of approach, which is called coordinate retrieval, the machine does not test for syntagmatic relationships in the answering of an encoded request for documents, but merely for the presence or absence of specified descriptors. The machine, in fact, is often a card-sorter rather than a computer. In the second type of classification, provision is made for syntagmatic relationships. This type provides a more reliable base upon which to set up a computerised semantics for mechanical translation and information retrieval.

3.2.1.1 What all coordinate retrieval systems have in common may be observed through a matrix. Ledley's[19] "Tabledex" provides one, in which the rows represent documents and their reference numbers, and the columns, the descriptors. An intersection of a row and a column (a posting) is assigned the digit, one, if a particular descriptor

does pertain to a certain document, and a zero, if it does not. A user interested in documents having to do with the application of nuclear theory, for example, would examine the columns of the descriptors application, nuclear and theory in figure 3, until all the rows in which the three postings of the descriptors containing ones are identified. These rows indicate the requisite documents.

For computational procedure the matrix has to be partitioned into entries. Each entry may consist of either a descriptor followed by references to documents to which it is pertinent or of a document number followed by the descriptors which appropriately describe it. Again the entry may consist of many descriptors followed by references to many documents or vice versa. In the first case the entries are of the type 'versatility: Powell, Tove' and 'analysis: Pope, Stockendal, Tove'. In a search for documents having to do with versatility analysis, Tove is retrieved by the matching of document references. The speed of computation is guaranteed by the alphabetical order of the descriptors.

In the second case the entries are of the type, 'counting, evaluation, versatility: Abrahams' and 'application, concept, design, England: Smith'. In this type of organisation multiple entries are often provided. For example, each of the entries is listed three times in the first example and four times in the second so as to bring each descriptor to the head of an entry. One of the alternative entries to that of the first example might be 'evaluation, counting, versatility: Abrahams'. With these multiple entries, the

		1.1 hysteresis	1.2 mass	2.1 counting	2.2 gas	2.3 thermal	3.1 differentiation	3.2 versatility	4.1 analysis	4.2 Netherlands	4.3 nuclear	4.4 theory	5.1 application	5.2 England	5.3 instrumentation	6.1 evaluation	6.2 concept	7.1 design
1.1	Abrahams	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
1.2	Aravindakshan	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
1.3	Gasstrom	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0
1.4	NBS Circular	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1
2.1	Nicholls	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
2.2	Pope	0	1	0	1	0	1	0	1	0	0	0	1	1	0	1	0	1
2.3	Powell	0	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	1
2.4	Seidle	0	0	1	0	0	0	0	0	1	1	1	1	0	1	0	1	0
3.1	Senior	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
3.2	Smith	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1
3.3	Stockendal	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	0
3.4	Tove	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0

Figure 3: Ledley's Tabledex.

identification of documents entered into an author-code dictionary is aided by systematic organisation. Since every descriptor is now at the head of some entry in the dictionary, alphabetical order may be utilised, although this convenience is at the expense of an increase in the number of entries.

An example of the third type of entry is to be found in Mooers'[20] Zatocoding. The representation of the descriptors is accomplished by assigning to them certain numbers, which are notched on to the top of a card. For example, where the descriptors 'selective device' (3,11,15,39) and 'film tally' (14,17,22,30) appear on the same entry card, the numbers 3,11,14,15,22,30, and 39 are notched in this order. Below the notched portion each descriptor is printed with its document references as shown in figure 4. Since the machine is searching for numbers specifically rather than for descriptors and document references, the numbers are set up carefully. If there had been other descriptors 'photographs' (14,11,15,39) and 'film production' (3,17,22,30) both represented by notches on this same card, the above notched numbers would be ambiguous because many descriptors might be retrieved through the same numbers.

3.2.1.2 The functioning of coordinate retrieval methods depends on an appeal to practical criteria, namely the limited number of ways in which words are in practice construed in natural language. Mooers[21] says: "In analysis we make no attempt to take the message of a document and to write a little abstract using descriptor words in such a way that the message of the document is preserved.... At the

Descriptors	Zatocodes	
selective device	3 11 15 39	U.
film tally	14 17 22 30	Ra
photo-electric sensing	1 11 34 40	Ri
audio frequency code	9 16 29 31	
camera	1 8 29 34	on
flash	17 23 34 38	
counting	8 26 33 37	

Figure 4: F

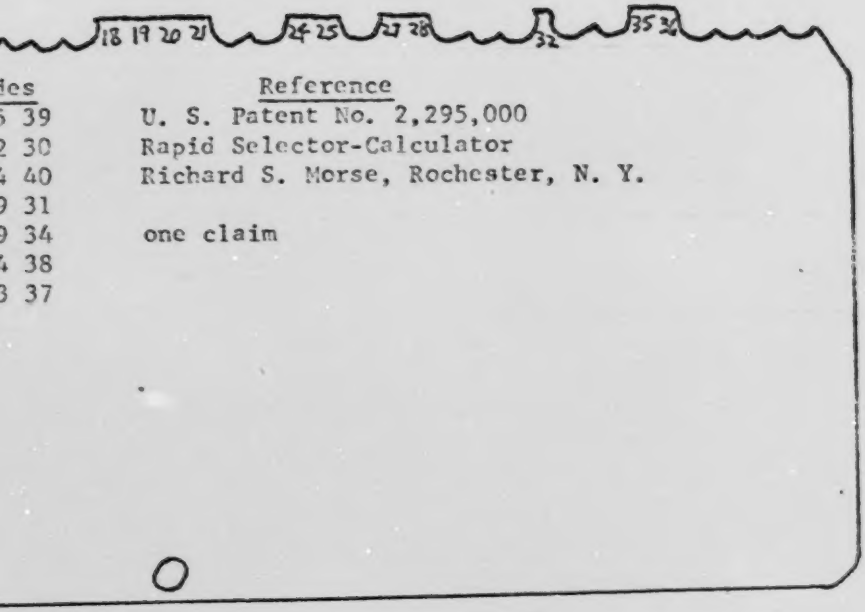


Figure 4: Format of Mooers' entry card.

symbolic and coding level retrieval, and not message preservation must be our goal". While the number of possible sentences in a language is infinite and therefore beyond the scope of a finite number of descriptors, out of which only a finite number of combinations may be formed, not all sentences are likely to occur in document titles.

Coordinate retrieval systems have been found to work well in the encoding of diagrams, in which the number of different components is limited. At the United States Patent Office[22] the diagrams of chemical structures provided ready-made descriptors in the form of the atoms in these structures. These were used in what the Patent Office group called the first topological system, in which each atom was assigned an identification number in terms of which a request was formulated. In mechanical procedure an atom to atom match was made between the encoded form of a user's request and each chemical structure for which there was a document and in the case of structural correspondence, the document reference was printed out. Since the diagrams were able to represent thousands of compounds, the first topological system was replaced by a second one, in which the number of descriptors was reduced by representing not atoms but groups of atoms in a compound set up according to the likelihood of their being requested.

The capacity of coordinate retrieval systems may be extended without an increase in the number of descriptors by the judicious assignment of many meanings to each descriptor. Gardin[23] found that

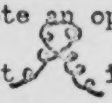
in archaeological drawings some components were predictable from the presence of others. For example, a man usually milks a ewe into a container, shears it with a tool and feeds it with fodder. For these situations the most appropriate descriptors would appear to be man, ewe, container, tool, fodder, milk, shear and feed. Since selectional restrictions limit the number of possible combinations of such descriptors, the principle of economy is taken into consideration. Gardin himself assigns many meanings to descriptors according to connotation. White represents non-pejorative acts including milk, shear and feed, and black pejorative ones like kill and strangle.

Brisch,[24] in his building classification, similarly utilises many meanings, although they are so apportioned to descriptors ("chapeaux" in this scheme) that the resolution of them requires a builder's intuition. The "chapeau" is a two-digit number with a broad range of meanings. The first digit represents a main class and the second a subclass - according to the principles of Dewey's decimal notation. The number 60, for example, refers to 'Functional components. Elements. Parts of buildings. Installations' and 62 refers to 'Foundations. Walls. Damp-proof courses. Partitions. Pillars. Arches. Dressings. Flashings'. This range of meanings is narrowed down by forming the combination 62-24-42 from two other "chapeaux", 24, 'Bituminous materials. Mastics. Lubricants. Fuels and Gases' and 42, 'Bricks. Blocks. Slabs. Slates. Tiles and Shingles'. The matching of meanings by someone with a builder's training would reveal that a document referenced by 62-24-42 referred to 'a bituminous, damp-proof course in a brick wall' or 'a bituminous

dressing on a brick wall'. The setting up of many meanings is intricate and could make the updating of a coordinate retrieval system more difficult than it has to be.

To overcome the obstacle of syntagmatic relationships in the setting of coordinate retrieval systems, stopgap measures have been suggested. Taube[25] points out that a document on 'fish as food' may be distinguished from one on 'food for fish' by assigning the descriptors food and fish to the first document and food, fish and plankton to the second. Role indicators[26] too have been advocated to discriminate between not only homonyms of the type base (alkali) and base (foundation), but also between the different roles that the same term may play. For example, 'lead as product' is assigned the descriptor lead I and 'lead as raw material', lead II. Role indicators may function like Latin endings to represent different syntactic functions.[27] For example, 'A man attacks a lion' would be represented by the three descriptors, man I, lion II and attacks, while 'A lion attacks a man' would be assigned the descriptors, man II, lion I and attacks. The distinguishing of role indicators by Roman numerals above is strictly mnemonic, a means by which a human may keep track of the descriptors. From the computational point of view it is sufficient that man I is different from man II just as it is different from attacks. While role indicators effectively remove ambiguity they do so at the expense of an increase in the number of descriptors.

3.2.1.3 The necessity of syntagmatic structure is a fact of not

just language but any notation with the capacity to convey an infinite amount of information through rules of grammar that allow the construction of syntagms that are necessarily infinitely long from a finite vocabulary. Circumstantial evidence may be found in a corollary of Gardin's[28] classification of ornaments. In his system a given decoration is assigned a symbol (called a radical) and the operations that it undergoes are denoted by affixes. For example, the decoration is named fix and a plurality of them are fixuli or ulifix. At this point the radical and affix appear to be unordered descriptors. However, fix may denote an operation as well as a particular decoration. The ornament  is called 'fix uli FIX' ('fix uli' in the shape of a fix). The repetition of fix, albeit capitalised in the second instance, is evidence of syntagmatic grouping.

While no ambiguity would result from scrambling the radicals and affixes in the above example, other examples may be found in which it would result. One may consider a decoration of the type , which will be named circ. With this notation the ornament would be labelled 'fix uli circ' (a group of fix in a circle) and , 'circ uli fix' (a group of circles in the shape of a fix). The two ornaments are represented by the same elements but in different groupings. In the first case fix and uli are immediate constituents, whereas in the second circ and uli are. Role indicators could be utilised to discriminate between the ornaments by means of unordered descriptors. With Gardin's[29] capitalisation constituting an indicator, the first ornament would be represented by the string of

elements 'fix uli CIRC' and the second, by the string 'FIX uli circ'. These elements are unordered but only because of an increase by two in the vocabulary of elements to include CIRC and FIX. To be able to represent all possible combinations of symbols and operations by such means that might occur in other subjects as well as archaeology, the vocabulary would have to contain an infinite number of elements. Alternatively, a machine might be programmed to abstract the meaning of a role indicator in isolation. Such a procedure, however, would involve syntagmatic structure. In 'fix uli CIRC', CIRC is one element on a higher linguistic level, but two elements on a lower level, namely circ and capitalisation.

The type of notation in which information is conveyed by pairs of items consisting of a term and a role indicator (Fillmore's[30] case) is a concession to the necessity of conveying syntagmatic structure. The term resembles Gardin's[31] radical, and the role indicator, the operation it undergoes. An example from chemistry, of an experiment in which hydrochloric acid and marble chips are mixed to produce carbon dioxide, may be considered. This situation may be represented as follows: (Hydrochloric acid, Agent 1) (Marble chips, Agent 2) (Carbon dioxide, Final product), in which the first component in each bracket is a term and the second, a role indicator. The pairs (liquid, Property), (solid, Property), and (gaseous, Property) might also be pertinent to the situation. The integration of them, however, into the rest of the notation would call for more linguistic levels as expressed by further parentheses as follows: (liquid, Property (Hydrochloric acid, Agent 1)) (solid, Property

(Marble chips, Agent 2)) (gaseous, Property (Carbon dioxide, Final Product)). The parentheses are required, since not only can individual terms and role indicators be immediate constituents, but also pairs of items.

In some amended versions of coordinate retrieval of the type used at the U.S. Patent Office, linguistic levels are indicated by interfixes rather than by brackets. Prior to the insertion of them the descriptors lead, copper, coatings and pipes pertaining to a request for documents on 'lead coatings for copper pipes' would each be assigned the same document number, 100.[32] While a document with the above title would be retrieved, one on 'copper coatings for lead pipes' might also be retrieved. Under the interfix system lead and coatings would each be assigned the document number 100A, and copper and pipes, the number 100B. The matching of the interfixes (A or B) would show the links between descriptors so as to discriminate between a request for '(lead coatings) (copper pipes)' and one for '(lead pipes) (copper coatings)'. Since in practice a user is often unable to retrieve what he wants, it is desirable to break a request into kernel parts, in this example to retrieve documents on 'lead coatings', 'copper pipes', 'coatings for copper' and 'coatings for pipes'. Since the meaning of for is needed to distinguish 'copper coatings' from 'coatings for copper', one must provide a more complicated set of interfixes than that of the U.S. Patent Office and Taube, in the following representation: lead 100A, coating 100A Q3 P3, for 100 Q3 P3, copper 100B Q3, pipes 100B P3. The link between copper, for and coating is shown by the interfix Q3, and the one

between pipes, for and coating, by the interfix P3. In each case the number designates the quantity of items linked.

The interfixes do not denote the type of relationship between terms but merely the presence of one. Even with their inclusion a document entitled 'The Destruction of Dyestuffs by Bacteria' would be encoded no differently from one entitled 'The Destruction of Bacteria by Dyestuffs'.[33] In each case the numbers would be of the type: 200AB Destruction, 200A Bacteria, 200B Dyestuffs, where 200 is the document reference and A and B are the interfixes. To discriminate between the two titles either of and by must be included as descriptors or the above descriptors must be assigned role indicators.

3.2.2 There are two methods of denoting the type of relationship existing between any two terms. One consists of fusing together analets (a construction consisting of a relationship item sandwiched between two terms). The other is a development of the role indicator method. An example of the first method is Farradane's[34] system, in which there are nine operators, as follows:

	Non-time-relation	Temporary relation	Fixed relation
Concurrent	Concurrence $\emptyset$	Comparison /*	Association /:
Not distinct	Equivalence $\neq$	Dimensional /+	Appurtenance /(
Distinct	Non-equivalence (Distinctness) $\neq$	Reaction /-	Causation /:

Figure 5: Farradane's Operators

To apply these relationship operators expressions not amenable to them such as 'The cat eats the mouse' are normalised into expressions of the type, 'The cat has food' and 'The food is a mouse', which are. Compound constructions consisting of many analets are incorporated by means of square brackets, the meaning of which may be observed by examining three synonymous analets. For the document title 'The Production of Glucose by Hydrolysis of Sucrose' the three possible analets are as follows: Glucose :/ Sucrose -/ Hydrolysis, Glucose :/ [Hydrolysis /-] Sucrose, and Sucrose [-/ Hydrolysis]/: Glucose. An operator immediately outside the brackets does not link with the term inside but with the one on the other side of them. These brackets provide for variety in the representation of the same title.

Each of the operators is comprised of two components. For example, '/:' (or ':/'), causation, consists of ':' stating what was caused and '/' stating what did the causing, and in '/-', reaction, '/' denotes the reactor and '-', what was reacted upon. In the role indicator method the title would be represented as follows: ((Glucose:) (Sucrose,/)) ((Sucrose, -) (Hydrolysis, /)) or, if interfixes A and B are used, as follows: (Glucose, :, A) (Sucrose, -, B) (Sucrose, /, A) (Hydrolysis, /, B). From a comparison of the two types of notation the operator is observed to be a concatenation of two role indicators, in which the terms to which they belong appear on either side.

3.2.2.1 The use of analets was favoured by Gardin[35] in SYNTOL (Syntagmatic Organisation Language), which he applied to information

retrieval in general. In his notation the analets may be expressed in terms of one dimension as Farradane's[36] were, or in terms of two as follows: $\begin{matrix} m \xrightarrow{Ri} p \\ Rj \downarrow \quad\quad \downarrow Ri \\ n \xrightarrow{} q \end{matrix}$, where m, n, p and q are terms and Ri and Rj are relationship items. This arrow format resembles that of Schank's[37] artificial intelligence, except that in SYNTOL utterances are normalised to preserve the analet. An utterance of the type 'a inhibits the effect of b on c' is not represented as $\begin{matrix} a \\ \downarrow Ri \\ b \longrightarrow c \end{matrix}$ but as $\begin{matrix} a \\ \swarrow Ri \\ b \longrightarrow c \\ Ri \end{matrix}$ which literally means 'a has an effect on b with respect to b's effect on c'.

There are four main categories of relationships and four of terms. The relationships are of the following kind, to which examples of contexts in which they occur are appended: R_1 predicative ('increasing.....unemployment'), R_2 associative ('cancerous... organs'), R_3 consecutive ('the effect of electricity...on muscles') and R_4 coordinative ('the differentiation of father and mother roles'). The categories of terms are as follows: Predicates, Entities (E1 or E2), States (S) and Actions (A). They may be observed in the representation of the utterance 'dégénérescence des muqueuses par un acide' or more specifically of its paraphrase 'action d'un acide sur la muqueuse: dégénérescence' $\underset{\text{acide}}{E2} \text{---} R3 \longrightarrow \underset{\text{muqueuse}}{E1} \text{---} R2 \longrightarrow \overset{S}{\text{dégénérescence}}$ The category, Action, is assigned to degeneration

when it occurs in phrases of the type 'acides dégénérants', in which it constitutes an intrinsic property of an acid.

While the four categories of terms and relationships may appear to be too few to prevent ambiguous representation in some circumstances, provision is made for expansion. For the utterances 'bénédiction de l'eau' and 'la phobie de l'angoisse' the diagrams 'bénédiction——R2——>eau' and 'phobie——R2——>angoisse' are inexplicit, because R2 does not differentiate the two meanings of de. To make the diagrams more precise, expansions are provided as follows: 'bénédiction——R2——>eau——R1——>Op instrumental' (bénédiction relates to water with respect to water being the instrument') and 'phobie——R2——>angoisse——R1——>Op. Signe' ('fear relates to anxiety as anxiety is a sign of it'). With the generation of these diagrams, a more appropriate name for R1 than predicative would be miscellaneous. While the main part of SYNTOL has to do with analets, the expansions involve role indicators, of which Op. instrumental and Op. Signe introduced by R1 are examples. The necessity for these expansions tests the usefulness of the main categories.

While the arrow diagrams represent the content of a discourse in detail, provision is made for an idioglossary type of summary. In Gardin's approach there are two main components. One of them, Source, provides bibliographic details about a document such as its date and the original language in which it was written. The other, Content, lists all the descriptors appropriate to the text of a document under seven headings, Scale, Theme, Focus, Beings, Space,

Time and Mode. In one example, that of a document on 'Lesions and Behaviour' the abstract was as follows: 'Modification of behaviour produced by lesions of the frontal and temporal lobes. Experiments on 36 monkeys and 10 cats. After temporal lesions, the animals are calm, cease to distinguish between edible and non-edible things. After frontal lesions they are apraxic and timid, a few manifest hyperactivity'. The headings and the terms listed under them for this abstract are as follows: 'Scale physio-psychology, Theme telencephalon attitudes, Focus 3, Beings cats, monkeys, Space , Time , Mode experiment'.

3.2.2.2 The role indicator method to which reference has been made in the previous paragraphs was applied at Western Reserve University. The role indicators (as described in Section 3.2.1.2) are analogous to Latin endings of the type a which identifies a noun like puella (girl) as the subject of a sentence. The analogy holds with respect to the constraints of linguistic levels (discussed in section 3.2.1.3) insofar as the endings allow free word order only within a given phrase or clause. At Western Reserve University,[38] a "telegraphic grammar" provided the means of representing various linguistic levels. In it, '¶' denoted the beginning or end of a paragraph, '&' of a sentence, '-' of a phrase and ',' of a subphrase. In addition, worksheets specified citation orders for the role indicators and their terms. In the encoding of the text of each document, neither parentheses nor interfixes were used to indicate grouping.

The telegraphic grammar consists of two parts, one for

analytic and the other for synthetic relationships. In the representation of the former, the role indicators take the form of infixes, which are based on an analogy with semitic languages. For the latter, the indicators are in the shape of separate three-letter combinations.

The infixes are added to semantic factors, which designate the most frequently applied concepts. A recognised method or technique like 'X-ray diffraction' or 'induction heating' being one would be counted as one multiple-word term to which a group of analytically related factors would be assigned.[39]

How the factors function may be observed in the construction of a code for the terms, tempering, stress relief and stress relieving.[40] The metallurgist, using his expert judgment, would provide for these terms the factors, M-TL (metal), P-SS (physical and chemical operations) and R-HT (processes and devices directly involving heat), in which the symbol '-' specifies that the appropriate infix is undetermined. The infixes are filled in by consulting a table. Since the metal is something acted upon, the infix 'W' (from the table) is inserted to give the complete factor, MWTL. Similarly PASS (in which 'A' is derived from the table) conveys the fact that tempering and stress relief are species of the genus, physical and chemical operations. RQHT denotes that the processes involving heat are the means to an end. The infixes having been determined, tempering, stress-relieving and stress relief are assigned the following respective groups of factors (which are unordered);

MWTL, PASS. RQHT. 013, MWTL. PASS. RQHT. 014, MWTL. PASS RQHT. 014. The capitalised factors designate the semantic content of the terms, 'Physical or chemical operations on metal by means of heat'. The numbers have to do with synonymy. When they match, the synonymy between two terms is complete, otherwise it is only partial.

In the second part of the telegraphic grammar, role indicators and terms are paired off, the former preceding the latter. The representation of the utterance 'in the analysis of Ni, Co ions interfere; Zn ions do not...' may be considered[41] which may be confused with the one for 'in the analysis of Ni, Zn ions interfere; Co ions do not...', for example, if syntagmatic links are ignored. The representation of it is as follows: -KEJ (material processed) Ni, -KAM (process) analysis, -KAM (process) interference, KQJ (by means of) ion, KUJ (component) Co, -KXM (process negation) interference, KQJ (by means of) ion, KUJ (component) Zn. The dash before each role indicator segments the representation into the kernel components, Ni, analysis, 'interference by means of Co ions' and 'non-interference by Zn ions'. This segmentation is a pragmatic measure. While the presence of more than one role indicator of the same kind might otherwise permit ambiguity within the representation as a whole, the segments are so constructed that in each one every indicator differs from every other one.

For retrieval, not only documents but also requests for documents have to be encoded. The encoding of a request is an exercise in marshalling encyclopedic data to specify what is implicit in it. In one for 'References to papers in which the electron band

theory has been applied to the study of beryllium',[42] the explicit key terms are beryllium (BQE) and 'electron band theory'. Further terms are derived through a knowledge of physics. A physicist would recognise that this request concerned energy (N-RG) and grouping (G-RP) with respect to the arrangement (R-NG) and location (L-CN) of the electrons (P-PH.6) of beryllium. Accordingly the formula for requesting documents is: BQE. P-PH.6. (G-RP.(N-RG+R-NG) +L-CN+N-RG). The notation works as follows: A,B means that the search is for documents characterised by the two descriptors A and B in this order, (A+B) that it is for those referenced by A and/or B, and (A.B) that only documents characterised by both A and B are requested. The above notation is reformulated request for documents concerning the location and/or energy of beryllium electrons or the energy and arrangement of groups of beryllium electrons. The reformulation consists of selecting documents that partly satisfy a request on the hypothesis that the full request cannot always be satisfied. While this notation is comprehensive, it was designed for the encoding of discourse by man rather than by machine.

The special classifications as a whole provide indices based on expert opinion by which to represent encyclopedic knowledge, but do not relate them to linguistic elements for a complete computational analysis. The Western Reserve group[43] note that between the words steel, hardening, quenching and hardness the relationships may be expressed in many ways in English: 'Steel is hardened by quenching', 'Quenching hardens steel' and 'Quenching produces hardness in steel'. However, no further inquiry is made.

Yet the categories are of interest in that they are the deep structure of which the three sentences are surface structures. From a cursory examination it appears that quenching may be tagged (technique), hardens (process, property), steel (material), produces (process) and hardness (property) to show how a notation for a discourse is derived from dictionary entries.

From an examination of the examples given in the special classifications, two principles emerge. While the Western Reserve group distinguishes analytic from synthetic relationships by format, the successful mechanical manipulation of data does not require such differentiation. A document title of the type 'Solubility of Bactericides Containing Mercury'[44] could be represented according to its paraphrase 'solubility of (things that destroy bacteria) containing mercury' as; (Solubility, properties given) (thing, properties given for) (bacteria, product) (destruction, process) (mercury, constituent), where the same brackets denote both analytic and synthetic relationships. The key to a computerised semantics lies in the uniform representation of both linguistic and encyclopedic knowledge. To encode 'Russia is making a survey of mineral resources'[45] one would not hesitate to adopt the more explicit paraphrase 'Some people in a government organisation in Russia are making a survey of mineral resources'.

Footnotes to Chapter 3

1. Vickery 1959
2. Vickery 1959
3. Bakewell 1968
4. Dewey 1965
5. Rangathan 1967
6. Bakewell 1968
7. Rangathan 1967
8. Rangathan 1967
9. Sayers 1967
10. Sayers 1967
11. Sayers 1967
12. Sayers 1967
13. Sayers 1967
14. Sharp 1965
15. Perry, Kent and Berry 1956
16. Rangathan 1967
17. Rangathan 1967
18. Micklesen 1961
19. Ledley 1959
20. Mooers 1955
21. Mooers 1955
22. Andrews et al. 1957
23. Gardin 1958
24. Brisch 1955
25. Taube 1954

26. Taube 1961

27. Gardin 1958

28. Gardin 1958

29. Gardin 1958

30. Fillmore 1968

31. Gardin 1953

32. Taube 1961

33. Vickery 1961

34. Farradane 1952

35. Gardin 1965

36. Farradane 1952

37. Schank 1969

38. Melton, Jessica 1958

39. Melton, Jessica 1958

40. Melton, John 1958

41. Melton, Jessica 1958

42. Melton and Perry 1958

43. Melton, Jessica 1958

44. Salton 1961

45. Rees 1958

4 GROUNDWORK FOR A COMPUTERISED DICTIONARY

4.1 Prolegomena to a Computerised Semantics

While the classifications elaborated upon in Chapter 3 provide semantic categories for the normalisation of natural language discourse, the task of translating from natural language into code and vice versa is left to man rather than machine. The notation for computationally construing dictionary entries and abstracting syntactic contexts step by step is not provided. Rangathan's[1] syntax of rounds and levels has significance in that it attempts to represent the linguistic levels according to which natural language is organised. Through the representation of the notation for his 'rice plant' example (Chapter 3, section 3.1.2) in the form of a tree, Rangathan's syntax is seen to resemble transformational grammar and Tesnière's[2] work which preceded it. Where the colon classification has five different facets, transformational grammar has parts of speech. The rounds and levels correspond to the N (Noun) P (Phrase) and V (Verb) P (Phrase) markers. However, Rangathan's syntax stops short of analysing specific syntactic contexts, which Su[3] and Noel[4] embark upon in recent research in computational linguistics.

4.1.1 Noel attempts to show the importance of semantic categories in a grammar, the lack of which has doomed various projects in mechanical translation to failure. To do so he adapts cases of the type advocated by Fillmore[5] to make them fit a syntagmatic tree, which he applies to compare the role of the conjunction and with that of

other function words. In many contexts this word acts only as a connective to link utterances of the type 'John crossed the road' and 'Mary climbed a tree' into a larger sentence like 'Mary climbed a tree and John corssed the road' or 'John crossed the road and Mary climbed a tree' without changing the meaning of the discourse. In some contexts, however, what precedes and is regarded as logically prior to what follows it. The conjunction is then an asymetric and. The sentence 'The machine was designed AND is used to search information' contains one and is represented in figure 6.

The nodes are of two kinds on this syntagmatic tree. One kind indicates the linguistic levels of an utterance and is labelled by an S followed by a number. At the S nodes a construction is designated +main insofar as it is a sentence, embedded or otherwise, and -main insofar as it represents a part of speech dependent on something else, a noun phrase, for example. At the top of the tree S zero represents the highest linguistic level, that of the discourse. S1 and S2 designate its constituent structures. Further subdivisions of either of these are denoted by S1' and S2'. The addition of primes to the symbol designates progressively lower linguistic levels. At this point the tree resembles immediate constituent analysis, where the descriptive main and its operators, plus and minus, specify the kind of dependency between constituents.

The second kind of node represents the cases and part of speech categories, verb, object, instrument, cause and goal. Prior to being parsed with these, the discourse is assigned an explicit

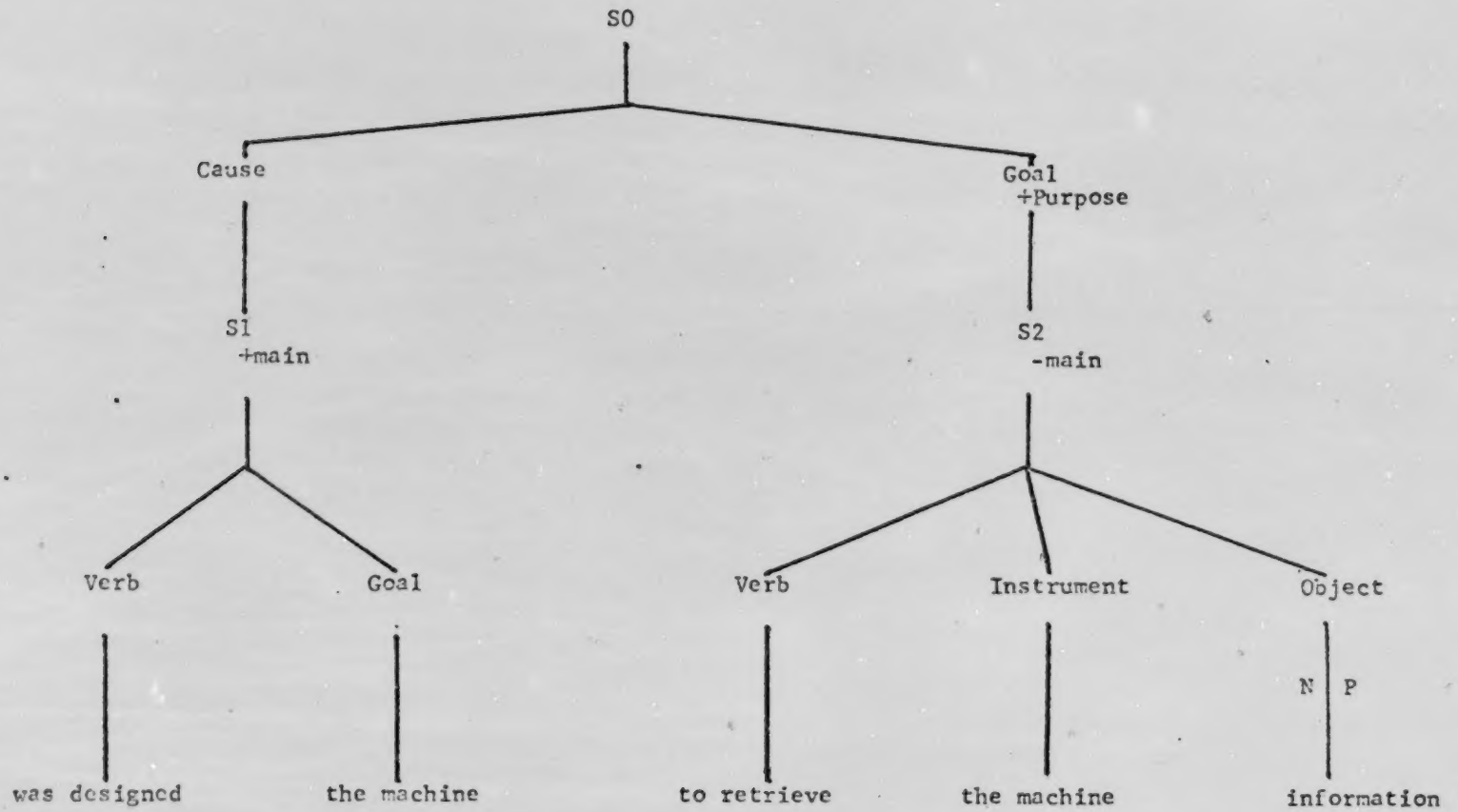
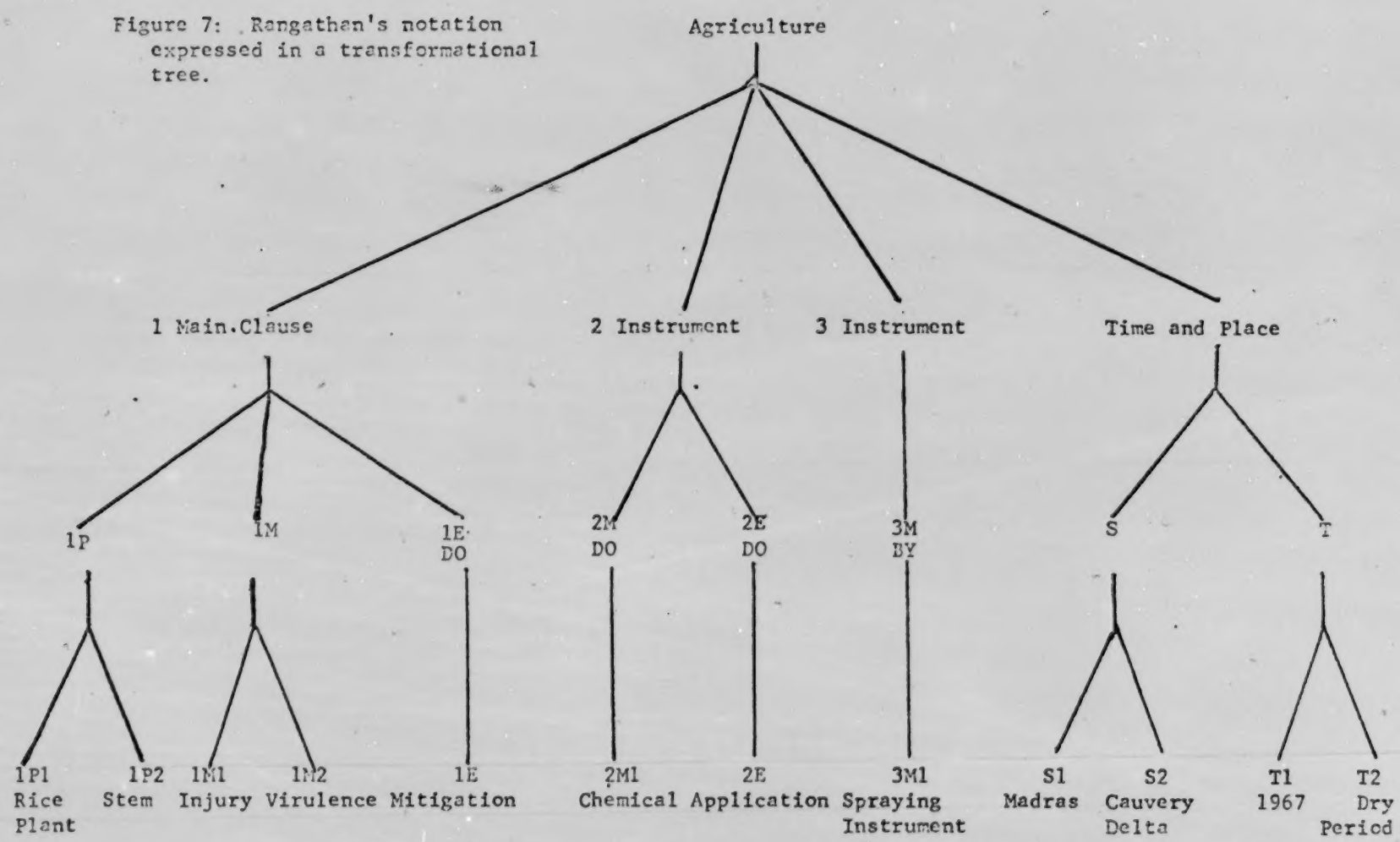


Figure 6: Noel's syntagmatic tree.

paraphrase to bring to light its kernel components. While at first the sequence of words at the bottom of the tree in figure 6 may appear inappropriate, it really is not when one considers it to be derived from the sentence 'Someone designed the machine; someone retrieved information by means of the machine'. How the cases function may be observed through the following utterances; (a) 'The center stores AND retrieves the stored information' (b) 'The storage of the information RESULTS in the retrieval of this (stored) information' (c) 'BY storing information, the center (is able to) retrieve this stored information' (d) 'The center stores information IN ORDER TO retrieve this stored information'. In each of these utterances there are two sentences whether embedded or not. The first states an action upon which the action of the second sentence depends, the retrieval of information being dependent upon its being stored. The first sentence in Noel's[6] terminology is designated as the Cause and the second as the Goal.

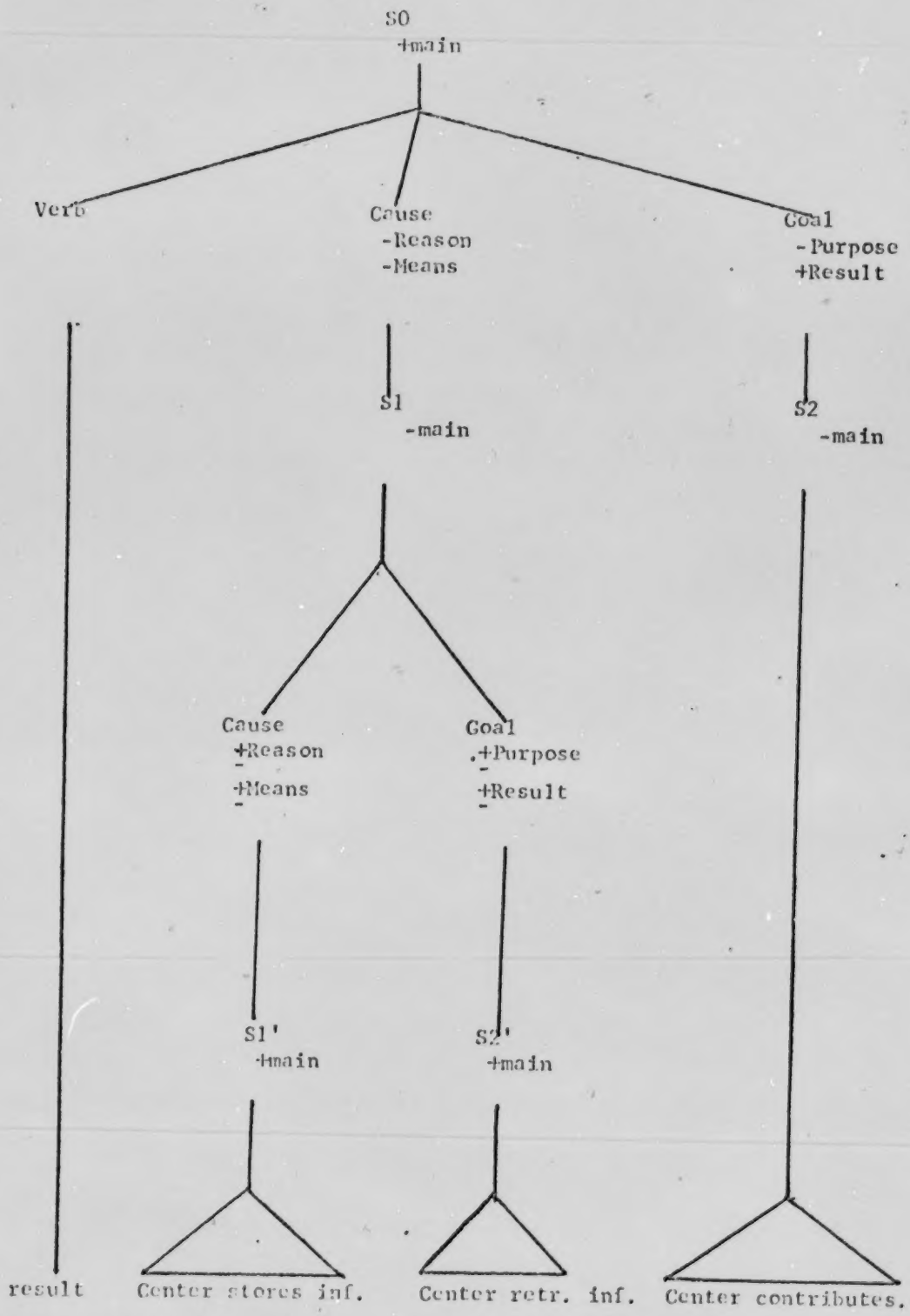
Differentiation of the above sentences is performed by the cases, Reason and Means, the species of Cause, and by Purpose and Result, the species of Goal, which all specify the presence or absence of human intervention. In sentence (a), since there is no specification of human intervention, the species cases are left unmarked. In sentence (b) the effect of storing information is considered to be outside human control; the first embedded sentence, then, is designated as -Reason, -Means and the second, -Purpose, +Result. In (c) the effect of storing information reflects human control but not necessarily human desire; the first embedded sentence

Figure 7: Rangathan's notation expressed in a transformational tree.



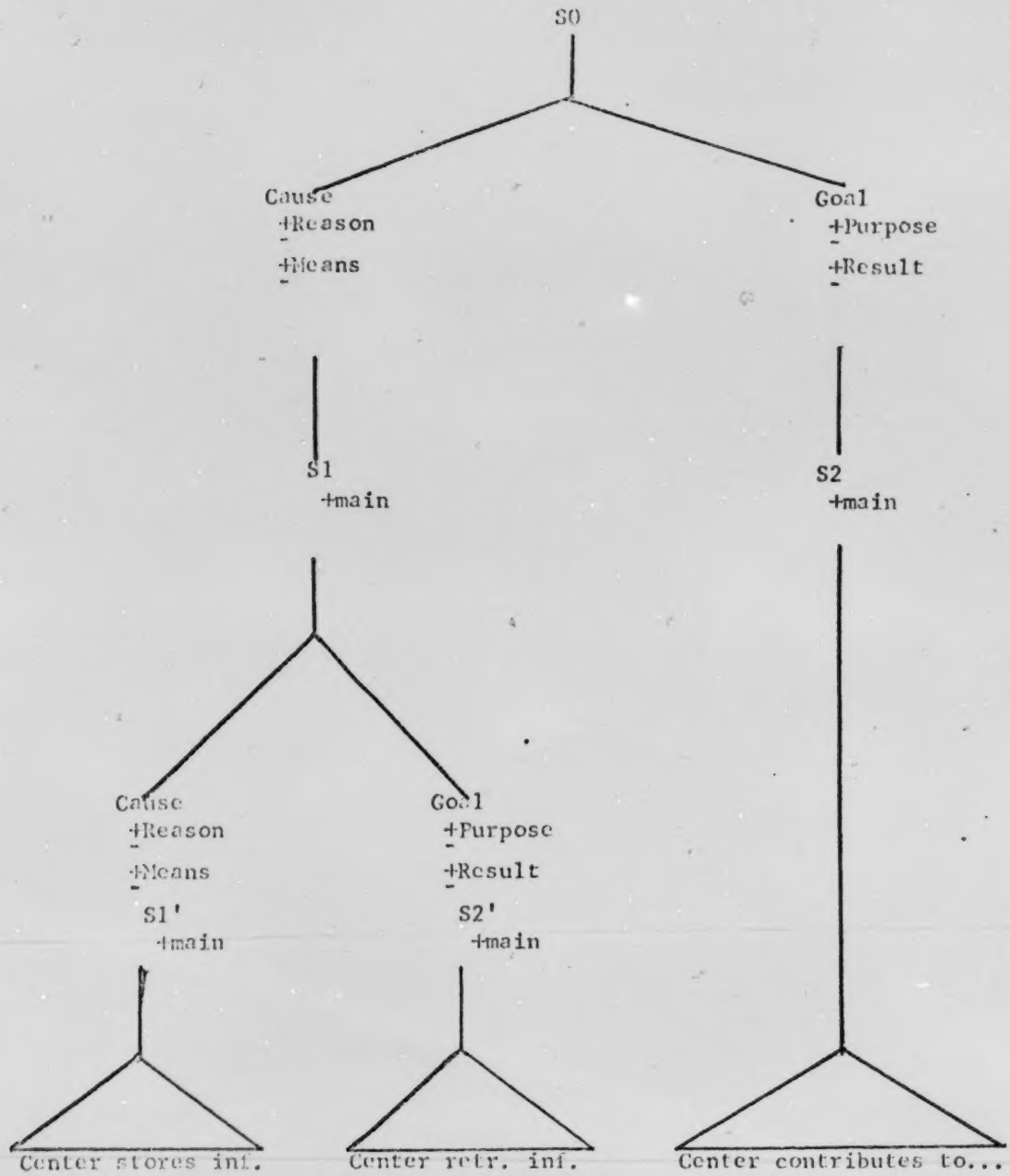
is correspondingly denoted by the cases, -Reason, +Means and the second by ±Purpose and ±Result. In sentence (d) the effect of storing is in human control and reflects human desire; the first embedded sentence, therefore, is designated +Reason, +Means and the second, +Purpose, -Result. In light of the above sentences asymmetric conjunction is observed to be a means of leaving the semantic content conveyed by the species cases unspecified. This neutralisation of specification may be visualised through the tree diagrams (shown in figure 8) of the sentences 'The Center's storage and retrieval of information results in the Center's contributing to the development of science' and 'The Center stores and retrieves information and contributes to the development of science.'

While Noel's cases have explanatory value, they need to be pared and supplemented. Where the cases +Reason, +Means, +Purpose, +Result and Instrument occur, corresponding categories of the type +Human +Wish, +Human +Control, +Human +Wish, -Control and -Human +Control could be substituted to provide more explicitness and greater economy in the inventory of cases. In order to implement Noel's case system the semantic categories of words surrounding and must be examined. While 'He stored and retrieved information' and 'He retrieved and stored information' are not interchangeable in Noel's contexts, the sentences 'He ate and baked the bread' and 'He baked and ate the bread' are, in the context of a preceding sentence 'Some people baked the bread and some ate it,' where the actions of the two verbs are considered to happen within one instant of time rather than consecutively.



'The Center's storage and retrieval of information results in the Center's contributing...'

Figure 8a: Noel's syntagmatic trees.



'The Center stores and retrieves information and contributes...'

Figure 8b: Noel's syntagmatic trees.

4.1.2 A notation for computationally construing dictionary entries for an unambiguous representation of natural language discourse is provided by Wilks[7] in his artificial intelligence. It is composed of semantic elements, which integrate the functions of descriptors and parts of speech. These elements form a dictionary definition of each of the possible translations ("stereotypes") of a word, thereby constituting part of a "formula". A formula embraces both semantic elements and a stereotype. For the English word red there are two French alternatives rouge and socialiste according to Wilks and therefore two formulae as follows:

(((WHERE SPREAD) KIND) (RED (ROUGE)))

((((WORLD CHANGE) WANT) MAN) (RED (SOCIALISTE))).

The last two items in each formula constitute a stereotype and the rest are semantic elements, of which the rightmost one indicates what part of speech to expect. Where the stereotype includes information about the target language's semantic and syntactic environment, it is called a "full stereotype". In the English-French entry for the word advise (as in 'to advise someone' and 'to advise someone to do something') there are two full stereotypes, in the first of which the meaning of FOLK as opposed to MAN is not apparent, as follows:

(ADVISE (CONSEILLER A (FN1 FOLK MAN)

(CONSEILLER (FN2 ACT STATE STUFF),

In the text, a grammatical sequence of words is often distinguished from an anomalous one by the fact that it matches one of a list of permissible syntagms of semantic elements called bare templates. For example, the grammaticality of the sentence 'John

(MAN) owns (HAVE) a car (THING)', in which the capital letters and brackets denote semantic elements, is registered by its matching the bare template, MAN HAVE THING. The identification of the template enables the computer to supply two-way dependency links for this sentence as follows: John<—>owns<—>car. An expansion of the above sentence like 'John owns a large (KIND) car' is represented by the addition of a one-way link to provide the full template, John<—>owns<—>car<— a. These two types of link constitute Schank's[8] conceptual structure. The representation of zero indices in a sentence is provided by "dummy" (D) semantic elements. For example, in 'John (MAN) talked (TELL DTHIS)' DTHIS denotes the zero presence of an object, and therefore that the verb is intransitive.

large

Bare templates are the criteria not only for establishing syntagmatic links but also for resolving ambiguities (which pertain to the individual word) and amphibologies (which have to do with a whole construction). Potentially the sentence 'This green bicycle (THING) is (BE) a winner (MAN? THING?)' may be parsed as THING BE MAN or THING BE THING. Since only the second parse matches a template, it is identified as the correct one. Where more than one template is applicable to an utterance, as in the amphibology 'They are eating apples,' which the templates MAN DO THING and THING BE THING fit, surrounding utterances will have to be searched, in this sentence, to identify the nearest antecedent of they.

The criteria for partitioning a text into portions for template matching are punctuation marks, subjunctions, conjunctions

and prepositions or a keyword. One such keyword is of, which in the sentence 'He has a book of mine' marks off a portion suitable for matching the template MAN HAVE THING. Where a word functions as a keyword only in some contexts, an absence of suitable templates to match the portions of a text will indicate an incorrect segmentation. That the partition of 'He (MAN) gave (DO) up (PDO) his post', in which up is a keyword, is incorrect is registered by the lack of a template corresponding to MAN DO PDO.

Dependency links between portions of a text are made through the keywords and what are called marks. In the sentence '(He came home) (from the war),' in which the brackets indicate partitions made by computer, the keyword is from and the mark to which it is linked is came. The detection by computer of links between portions is essential for the translation of ambiguous prepositions, the resolution of which requires the context beyond the prepositional phrase itself. For the preposition "out of" there are three alternative translations into French, which may be observed in the following sentences: (1) 'It was made (PR MARK *DO) out of (PR CASE SOURCE) (de) wood (FN1 STUFF THING)', (2) 'He killed (PR MARK *DO) him out of (PR CASE SOURCE) (par) hatred (FN2 FEEL)' and (3) 'I live out of (PR CASE LOCA) (en dehors de) town (FN1 POINT SPREAD)'. The capitalised items represent semantic elements that either correspond to the dictionary entry of each word or are established in the course of analysis and PR and LOCA are abbreviations for prepositional and location. The full dictionary entry for 'out of' consists of the following three stereotypes:

((PR CASE SOURCE) (PRMARK *DO) DE (FN1 STUFF THING))

((PRCASE SOURCE) (PRMARK *DO) PAR (FN2 FEEL))

((PRCASE LOCA) EN DEHORS DE (FN1 POINT SPREAD)).

The cases have to do with the preposition itself. The rest of the information in the stereotype is a statement of the syntactic and semantic environment of the preposition which is required for a given French translation to be assigned.

4.2 Structure of a Comprehensive Computerised Dictionary

4.2.1 Wilks'[9] mechanical intelligence as an effective means of integrating syntactic and semantic categories provides the foundation for a computerised semantics. It may be built upon by the addition of Katz and Fodor's[10] marker theory and by the idioglossary (notional family) approach to represent linked syntagms on the nodes of a tree. While the content of a dictionary entry is finite and therefore does not require a notational syntax considered necessary in chapter 3 for the encoding of a text, Katz and Fodor's unordered markers are on a precarious footing. They are parts of dictionary definitions between which the syntagmatic links have been severed. The consequence of such severing is pointed out by Vickery.[11] He indicates that a notation of simple interfixes would be inadequate to differentiate the utterance 'the destruction of dyestuffs by bacteria' from 'the destruction of bacteria by dyestuffs' by linking dyestuffs and bacteria to destruction. Without indicating the dependency links between these words as shown by of and by in natural language the notation can only go so far as to represent the utterance 'destructive

relationships between dyestuffs and bacteria'. The absence of the links marks the sentence as the genus of the previous two utterances, just as the absence of a symbol, in the following case DO, marks the word animal AN MA as the genus of dog AN MA DO in Perry's[12] code. A limited basis for unordered elements lies in the integration into the dictionary structure of idioglossaries, the use of which does not require manipulation of the syntax of a text.

The idioglossary approach capitalises on the fact that a word central to the topic of a text tends to be used consistently in the same sense so that where the local context of an ambiguous word fails to resolve it, a likely meaning may be assigned in default. In the sentence "Gray and his collaborators concluded that the suckers were acting as time signals for each phase of the movement, and are normal but not essential channels for peripheral excitation"[13] that sucker refers to an animal's anatomy in the context is determined by the topic, annelids. As a relatively simple approach, the idioglossary is worth incorporating into a semantic tree. Ambiguous words occur, however, that cannot be treated by the method. While sucker will refer to fish in a text on fish in most local contexts, there will be exceptions, which may be observed in the following quotations: "Sucker, a freshwater fish with thick soft lips that form a suckerlike mouth,"[14] and "Catostomidae (suckers) The suckers a family of fishes..., with the mouth so constructed that it can form a tubelike sucker."[15] Because of the necessity of examining the local contexts of words the idioglossary approach must be supplemented.

Despite the faults of Katz and Fodor's tree of markers, the structuring of definitions by means of it to show how a word's different meanings are related is a starting point for a more comprehensive tree. Since such a tree is subject to the criticism made in chapter 3, section 3.1.2, of all trees, that they cannot simultaneously represent all dimensions of knowledge, the searching of descriptors for matching purposes will not necessarily be in the convenient form envisaged by Katz and Fodor, namely, the branch by branch construing of information from the top of a tree to the bottom. For example, in a sentence in which the word sucker (figure 11) occurs, components that match N-thing-ER in terminal 2 may be identified before ones matching BY DO suction. Since the tree represents an entry in an English-Code dictionary and not a Code-English one and thereby corresponds to the author-descriptors entry of coordinate retrieval and not the descriptor-authors one of coordinate retrieval, it has all the difficulties of organisation that the former type of entry was claimed to have (chapter 3, section 3.2.1.1). Since some form of organisation is preferable to none, the tree will nonetheless be maintained.

To construct the comprehensive tree, modifications have been made of Katz and Fodor's tree. The residue of information that they put in the form of distinguishers at the bottom of the tree has been structured on to the tree itself. The bottom of the tree now contains stereotypes or full stereotypes, which will be called terminals. One might consider retaining the plus and minus indicators and the descriptives of the marker tree within the framework of the new tree

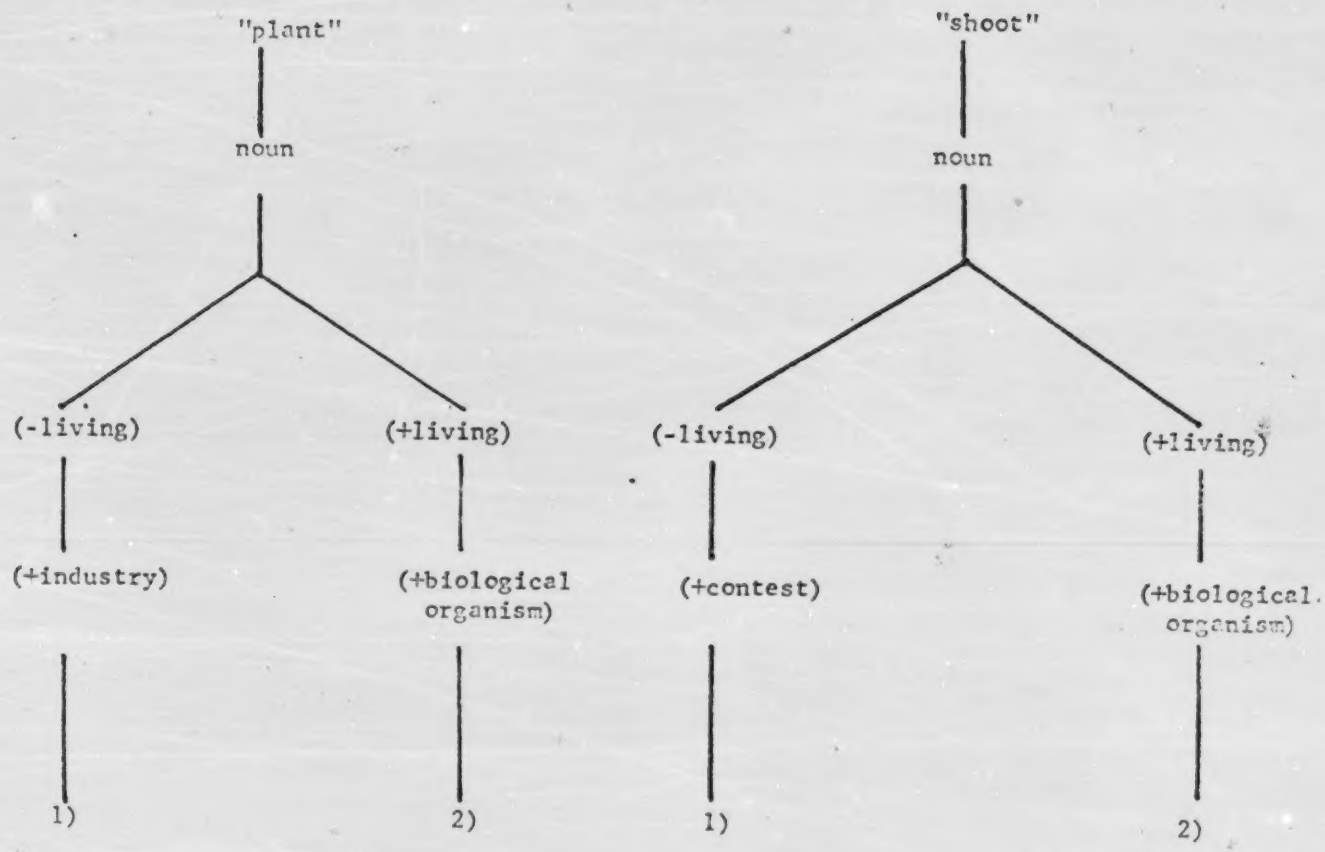


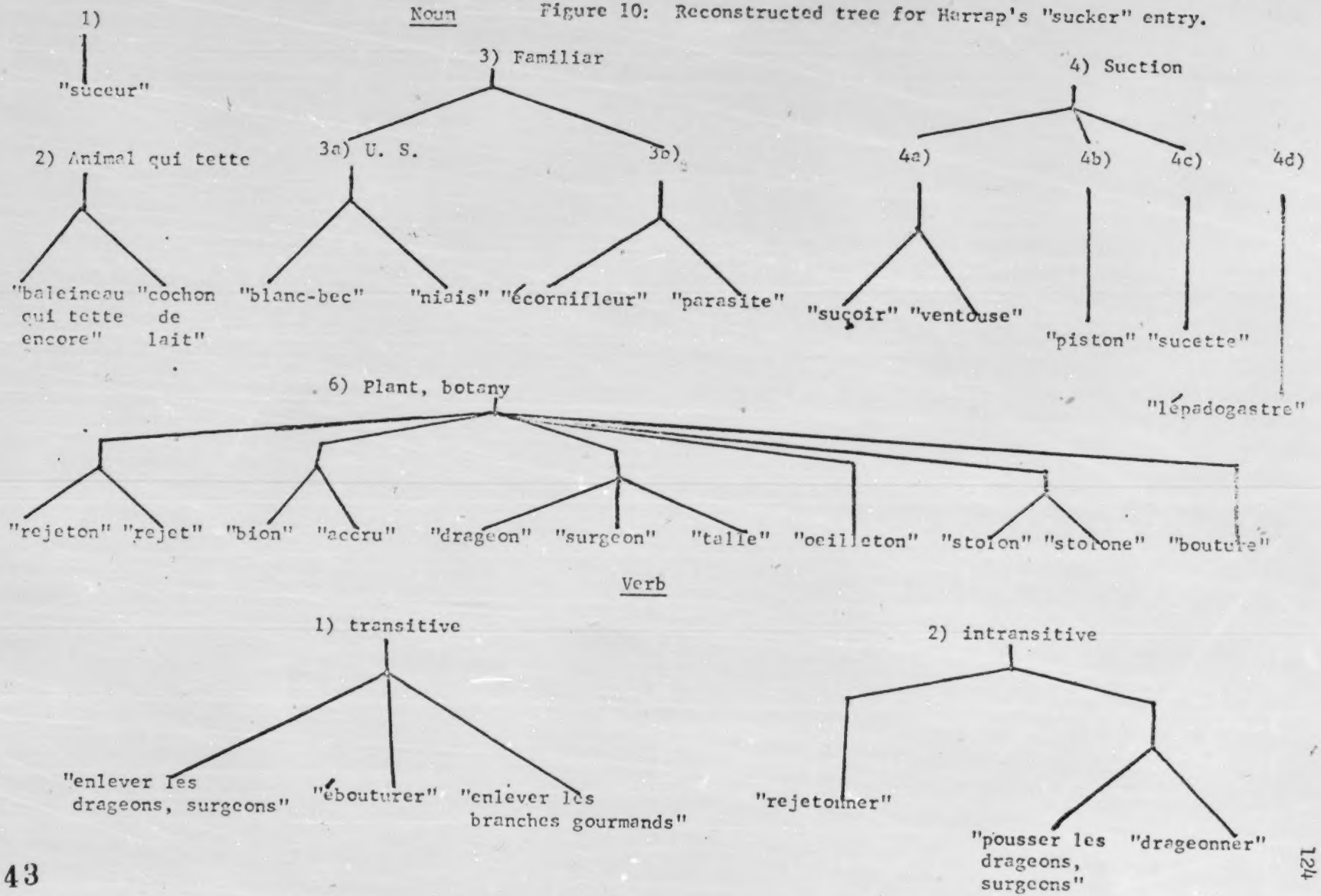
Figure 9: Binary branching of marker trees.

as some of many semantic elements. However, the limitation of the branching at the nodes to two, imposed by the indicators, and the consequent inflexibility in the setting up of categories would obstruct the purpose of the tree. Partial trees for the words plant and shoot may be considered. The binary branching into plus and minus categories in diagram 9 is consistent with what Katz and Fodor envisaged, but does not provide functioning trees. Matching of markers in the sentence 'The plant's shoots expanded' does not reveal the contextual terminals of plant and shoot, because both the marker (+living) and (-living) of each word finds a counterpart in the other. The defect lies in the indiscriminate application of the minus operator. The marker (-living) while symmetrical with (+living) on the alternative branch is vague and needs to be divided into more specific ones to prevent terminal 1 of shoot and plant from being selected.

With emendations, Katz and Fodor's tree resembles a conventional dictionary, in which the numbering of the definitions constitutes an organisation of them into semantic categories. The setting up of a tree for the word sucker as defined in Harrap's French dictionary[16] reveals that the definition numbers are in fact encoded markers. As a bi-product of trees constructed for computation, a dictionary would emerge in which there would be sufficient control of the wording of the definitions for relationships between words as well as meanings to be displayed.

4.2.2 To demonstrate the operation of the emended version of Katz

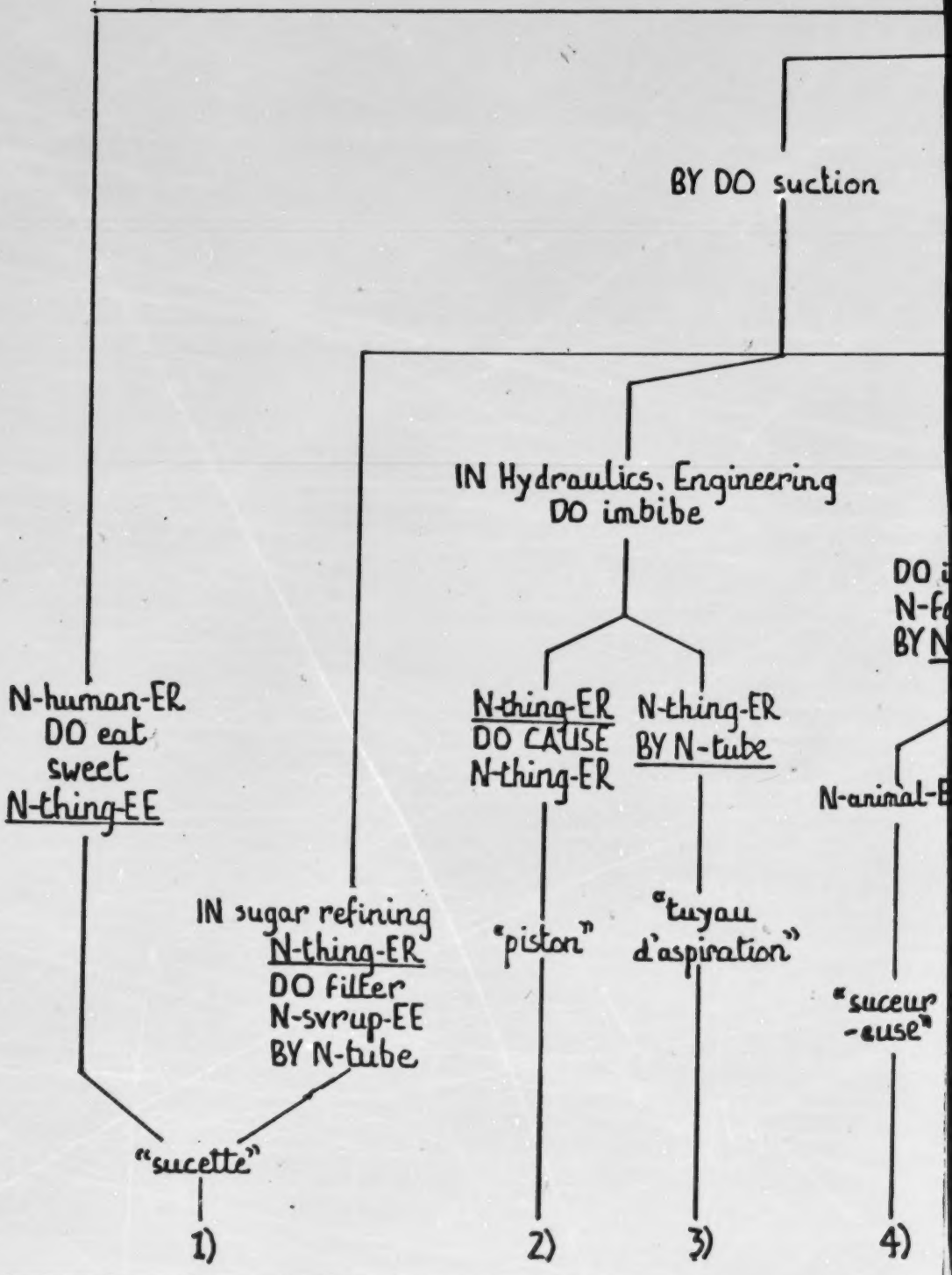
Figure 10: Reconstructed tree for Harrap's "sucker" entry.

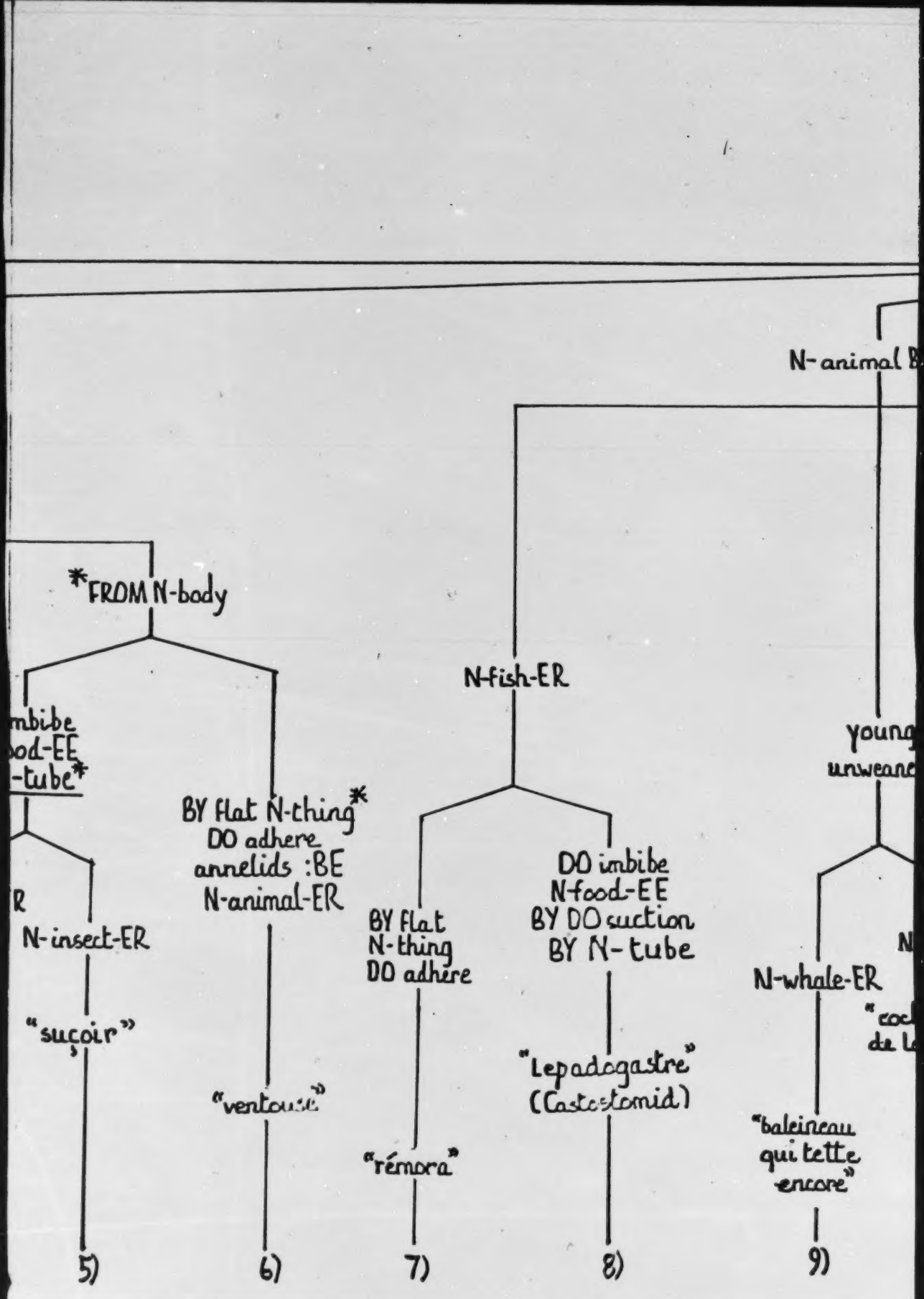


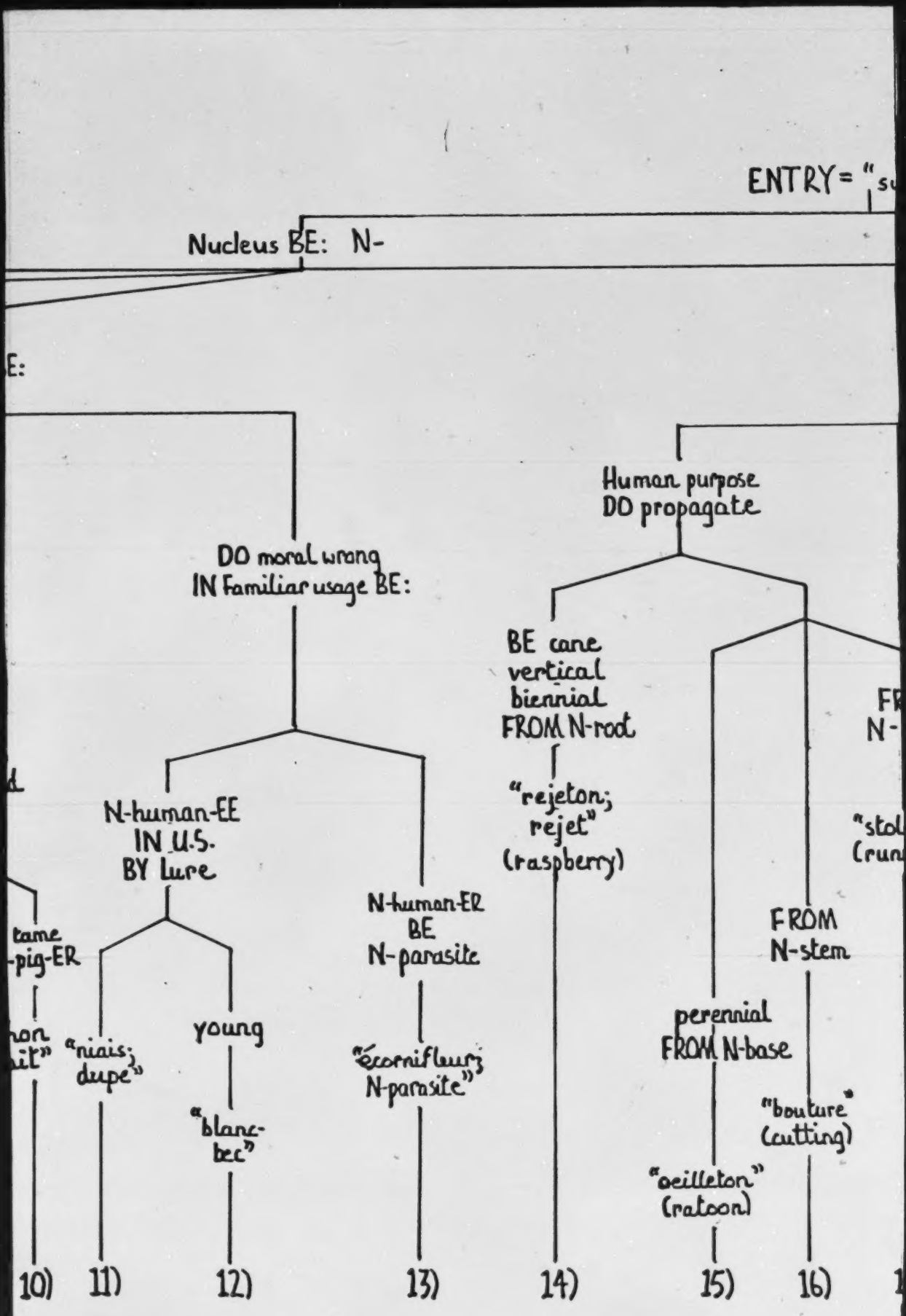
and Fodor's tree, which includes Wilks'[17] context specifying descriptors, the ambiguous words sucker and baste have been selected. Comparable words are creeper, runner and spring. Semantic elements were derived for the above two words by examining actual usage in texts and encyclopedias. In addition, Webster's Seventh Collegiate dictionary was combed for words such as swim in 'The sucker swims' that are permitted by selectional restrictions to form a direct syntactic link with sucker. These words were sorted into semantic categories. The two sources of categories were then collated. In the setting up of them the requirements for disambiguation were found to diverge from those for other purposes of computation. To meet the former, the semantic element human was sufficient to identify the terminals of sucker having to do with moral wrong. To make the tree an all-purpose one, all information was encoded.

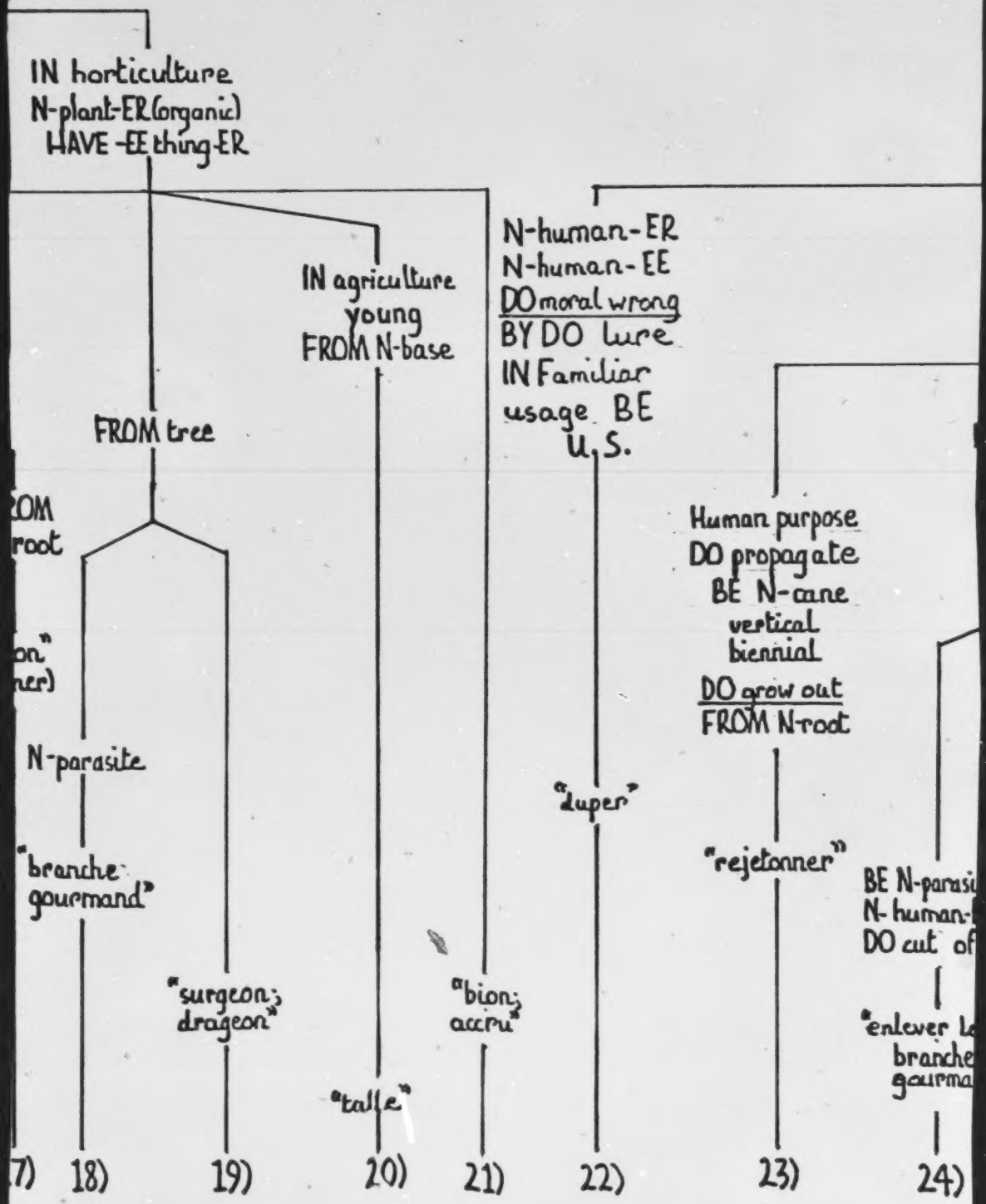
In the computational analysis of the local contexts of ambiguous words, the simplest operation consists of testing immediate constituents such as an adjective and a noun, a subject and a verb and a verb and its object and has been seized upon by computational linguists. Booth, Brandwood and Cleave's[18] (Chapter 1, section 1.3.2.1.1) and Masterman's[19] (Chapter 1, section 1.2.2.2) concept numbers are essentially a list of interfixes stating which pairs of words may become immediate constituents without forming anomalous constructions. The interfixes are like Katz and Fodor's markers but are not factored into semantic categories. The concept number technique is plausible in that the number of interfixes, though large, will be finite since only pairs of words are linked, but it has

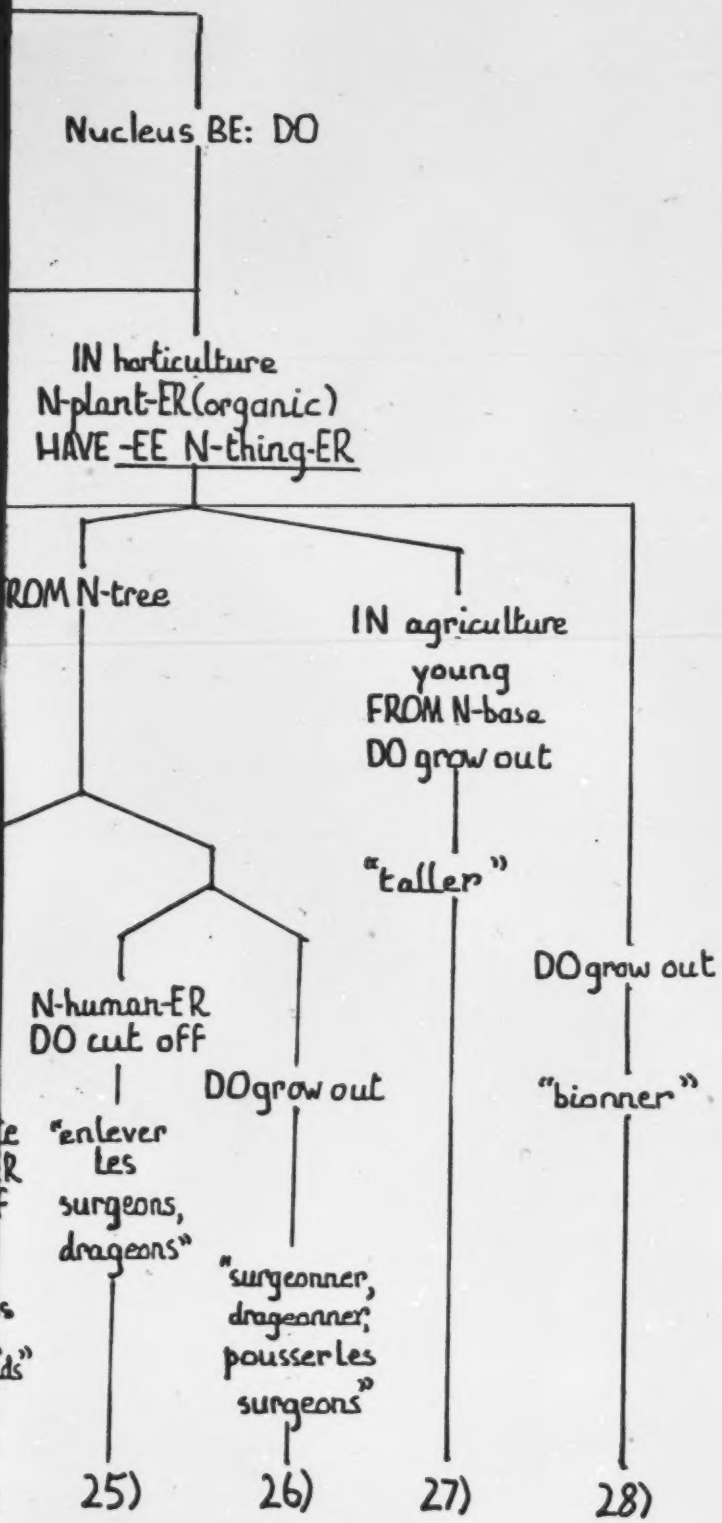
Figure 11: Model tree for "sucker"




N-animal B
*FROM N-body
N-fish-ER
young
unweane
mbibe
od-EE
-tube*
R
N-insect-ER
"suçoir"
5)
BY Flat N-thing*
DO adhere
annelids :BE
N-animal-ER
"ventouse"
6)
BY Flat
N-thing
DO adhere
"rémora"
7)
DO imbibe
N-food-EE
BY DO suction
BY N-tube
"Lepadogastre"
(Castostomid)
8)
N-whale-ER
"baleineau
qui tette
encore"
9)


ENTRY = "s
Nucleus BE: N-
E:
DO moral wrong
IN Familiar usage BE:
Human purpose
DO propagate
BE cane
vertical
biennial
FROM N-root
"rejeton;
rejet"
(raspberry)
FR
N-
"stol
(run
N-human-EE
IN U.S.
BY lure
N-human-ER
BE
N-parasite
FROM
N-stem
perennial
FROM N-base
"bouture"
(cutting)
"oeilleton"
(ratoon)
young
"blanc-
bec"
"niais;
dupe"
"écornifleur;
N-parasite"
10)
11)
12)
13)
14)
15)
16)


IN horticulture
N-plant-ER (organic)
HAVE -EE thing-ER
IN agriculture
young
FROM N-base
N-human-ER
N-human-EE
DO moral wrong
BY DO lure
IN Familiar
usage BE
U.S.
FROM tree
OM
root
on"
ner)
N-parasite
"branche
gourmand"
"surgeon;
drageon"
"talle"
"bion;
accru"
"luper"
Human purpose
DO propagate
BE N-cane
vertical
biennial
DO grow out
FROM N-root
"rejetonner"
BE N-parasi
N-human-
DO cut of
"enlever l
branche
gourma
7)
18)
19)
20)
21)
22)
23)
24)


Nucleus BE: DO
IN horticulture
N-plant-ER(organic)
HAVE -EE N-thing-ER
ROM N-tree
IN agriculture
young
FROM N-base
DO grow out
"taller"
DO grow out
"bionner"
N-human-ER
DO cut off
DO grow out
"enlever
les
surgeons,
drageons"
"surgeonner,
drageonner,
pousser les
surgeons"
25)
26)
27)
28)

limitations. In the sentence 'I see suckers', for example, see does not disambiguate sucker. Therefore, the limitation of the trees for sucker and baste to application within a single sentence is but a short-term expedient based on the insuperable difficulty of encoding all encyclopedic knowledge.

To organise semantic categories in terms of syntactic structure, semantic elements, including what will be called linguistic descriptors, have been mapped on to a tree to show what the definitions of different terminals have in common, without destroying the syntagmatic links between the components of each definition. For example, where one terminal of the word sucker is defined as 'a young unweaned whale' and another as 'a young unweaned tame pig', the two definitions are synthesised as follows: young unweaned N-animal BE—<tame N-pig / N-whale in which the first four components are represented on the upper part of the tree in figure 11.

The linguistic descriptors are a refinement of the traditional parts of speech for parsing, and thereby cover the area to which Bar-Hillel[20] applied his categorial grammar. In the tree diagram for sucker, that follows, linguistic descriptors are designated by capital letters. Their functioning may be observed in the words employer, employ and employee, from the suffixes of which some of the descriptors have been derived. In the usual paraphrases of them different words would be used in each case, as follows: 'one who employs someone', employ and 'one who is employed by someone'. Insofar as a word's semantic and syntactic environment

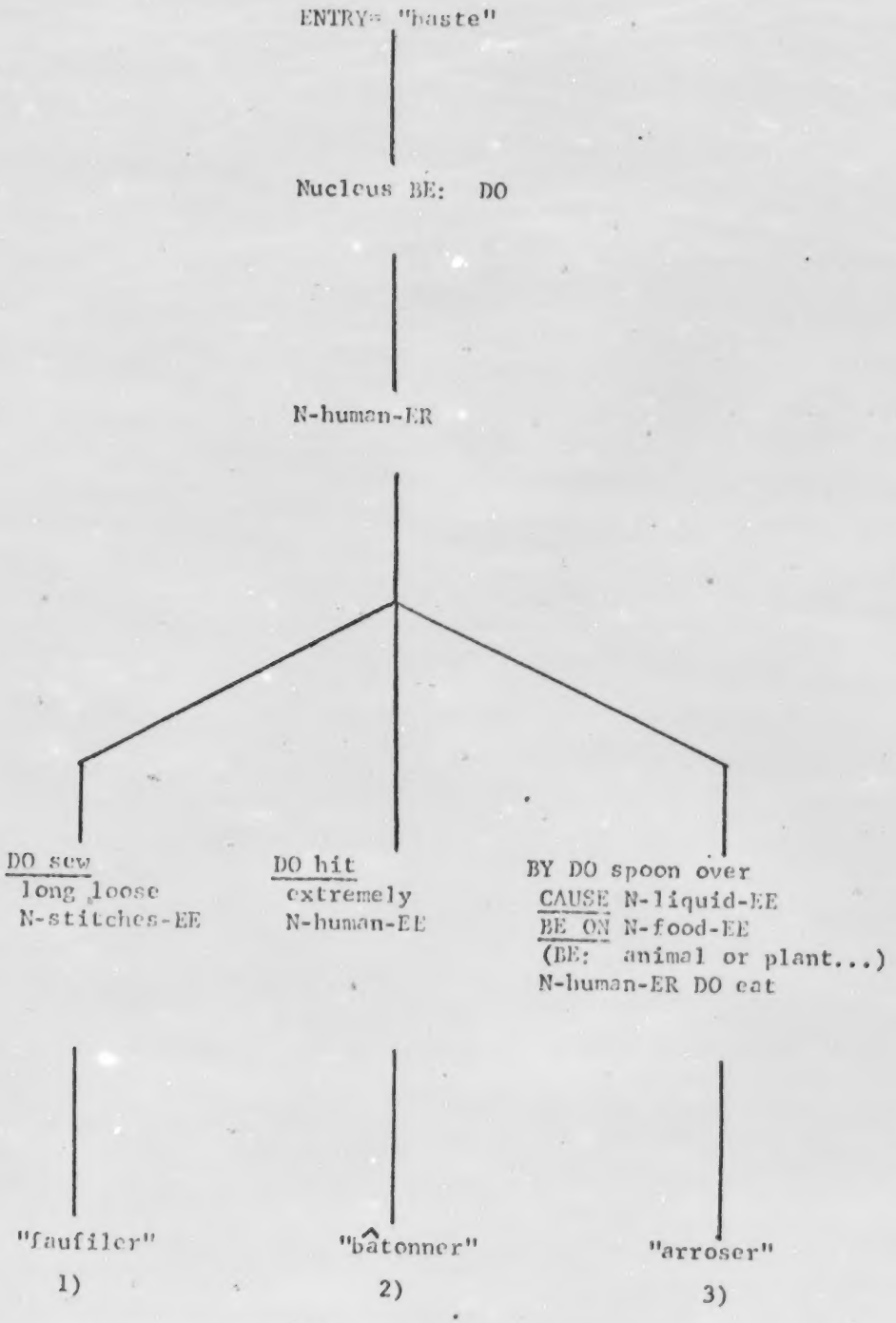


Figure 12: Model tree for "baste"

contributes to its meaning just as much as its paraphrase does, the three words merit the same notation 'N-human-ER DO-employ N-human-EE', in which 'N-' means noun, '-ER' subject, '-EE' object and 'DO-' verb. Insofar as the words belong to different parts of speech they are underlined differently in their definitions, as follows: 'N-human-ER DO-employ N-human-EE', 'N-human-ER DO-employ N-human-EE' and 'N-human-ER DO-employ N-human-EE'. The underlined portion is the nucleus of a word's definition and the rest is its environment, a means of crossreferencing it to the other two words.

In order to incorporate the idioglossary technique the tree diagram includes the linguistic descriptor, IN. It is through it that the hierarchy of idioglossaries elaborated upon by Micklosen[21] (Chapter 1, section 1.3.2.1.3) is represented. In the tree, hyponymy relationships, whether in the environment of IN or of some other linguistic descriptor, are denoted by BE: or :BE. The element nearest the colon is the species and the one furthest away is the genus. Whether BE: is used or :BE depends on how the tree branches and therefore, where the semantic elements are placed.

4.2.3 By means of the trees for the two ambiguous words sucker (figure 11) and baste (figure 12), it is possible to disambiguate one word by means of the other. In the analysis of the sentence 'He basted the sucker before eating it' the environment of DO sew at terminal 1 of baste is matched with each nucleus of sucker for elements corresponding to 'long loose stitches'. Since there are none, terminal 2 is inspected and the common nucleus of terminals 11,

12 and 13 of sucker, N-human, is found to meet the requirement of an object of baste (2). Terminal 3 of baste is similarly tested and the common nucleus of terminals 7, 8, 9 and 10 of sucker is found to qualify as an object of baste (3). To decide between terminals 2 and 3 of baste, the phrase 'before eating it' is examined, in which it is traced to sucker. The word it in its context narrows down the nucleus of sucker to N-animal, and therefore to terminals 7, 8, 9 and 10. As the object of baste belongs to the category, animal, terminal 3 of this word is selected as its contextual meaning.

Further disambiguation of sucker must wait for a wider context than the single sentence given and a more comprehensive tree based upon a lot of empirical evidence. Nonetheless, the one given illustrates the interaction of semantic and syntactic categories. The linguistic descriptors allow syntactic clues to pinpoint terminals. In the sentence 'The sucker basted the meat', the fact that sucker is the subject enables it to be categorised as human to limit the applicable terminals to 11, 12 and 13.

Since the definitions of terminals encoded on a tree are represented in deep structure, sentences may have to undergo language normalisation before the tree is usable. In order to pinpoint terminal 2 as the most probable meaning of sucker in 'The sucker drew the water up', this sentence must be assigned the paraphrase 'something (N-thing-ER) caused (DO CAUSE) the water (N-thing-EE BE N-thing-ER) to rise (DO) by (BY) the suction (DO) of the sucker (N-thing-ER)'. While the sentence does not perfectly match the

encoded definition for terminal 2, N-thing-ER DO CAUSE N-thing-EE BE N-thing-ER DO imbibe BY DO suction, it matches this one more closely than the ones for the other terminals of sucker.

The matching of a terminal's definition with a sentence shows how the contextual meaning of a word is derived not only from a selection of its dictionary alternatives, but also from the contextual meaning of another word or words. For the sentence 'He fooled the sucker with the stuffed tiger', the matching process shows how the stuffed tiger comes to be viewed as a lure. The stuffed tiger is so considered because of its link with a component of the definition of sucker (11) in the following mapped-out version of the above sentence: 'He (N-human-ER) fooled (DO moral wrong) the sucker (N-human-EE) with (BY) the stuffed tiger (Lure)'.

Further developments on the tree presented so far may include the incorporation of language normalisation programmes. A tree accordingly equipped would be capable of taking into account the facts covered by Fillmore's[22] cases, by matching the elements of an utterance with a dictionary definition which in turn would provide an explicit paraphrase of the utterance. An analysis of his sentence 'A man (N-human-ER) moved (DO move) the rock (N-thing-EE)' 'The wind (N-thing-ER) moved (DO move) the rock (N-thing-EE)' and 'The rock (N-thing-ER) moved (DO move)', mentioned in Chapter 2, section 2.2.1, would centre on the word moved. That it does not have the same function throughout becomes apparent in the following respective paraphrases of the above sentences: 'A man caused the rock to move',

'The wind caused the rock to move' and 'The rock moved'. In order to show how to arrive at these paraphrases computationally, the tree in figure 13 is provided. To determine which branch is applicable to the contextual function of move, the matching procedure described in previous paragraphs is used. By this means the surface structure notation for 'A man moved the rock' is recognised as 'N-human-ER DO move N-thing-EE BE; rock'. The deep structure representation is arrived at by replacing non-bracketed elements by bracketed ones that follow an equals sign. For the above sentence it is; N-human-ER (Means) CAUSE N-thing-ER DO Move. The viability of the above procedure would depend upon how complex the relationship between a given surface and deep structure was.

Often the key to disambiguation lies in a series of syntactic links. The following extract[23] may be considered; "The working of a Newcomen engine is.....a very painful process..... When the pump descends, there is heard a plunge, a heavy sigh and a loud bump; then as it rises, and the sucker begins to act, there is heard a creak, a wheeze,.....". In this sample the disambiguation of sucker by means of its link with pump relies on tracing the link through it. Since indirect linking of this type may take many forms, reliance will have to be placed upon a documentary language of the kind developed at Stanford or one based upon Gardin's[24] SYNTOL to reveal explicitly the linking between elements in a text.

Such a documentary language would be applicable to Katz and Fodor's[25] sentence (chapter 1, section 1.3.2.2) 'I shot the man with

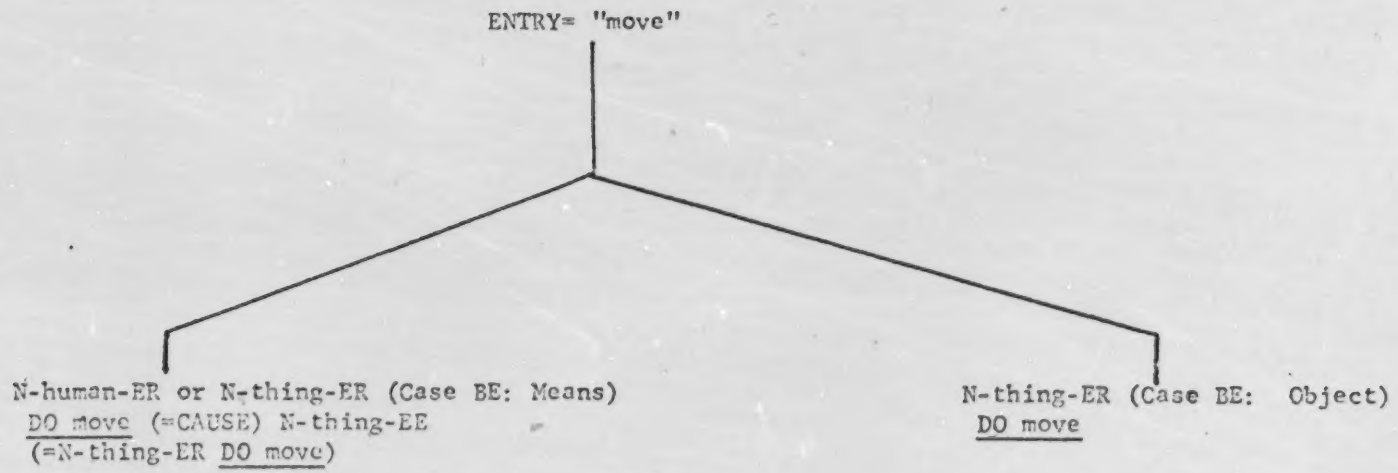


Figure 13: Deep and surface structure tree for "move"

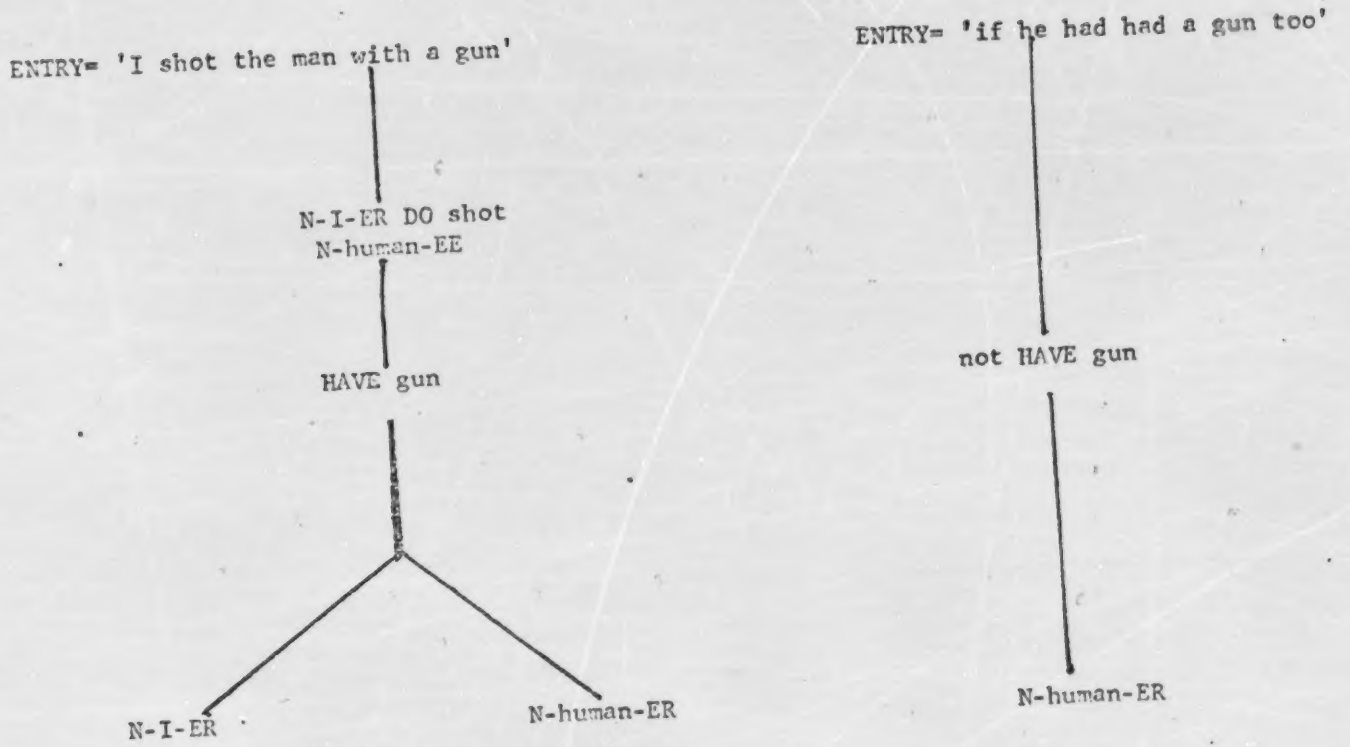


Figure 14: Computationally constructed tree

a gun, but if he had had a gun too, he would have shot me first'. A likely outcome of its use would probably be the components on trees in figure 14. In these trees the non-matching of 'N-human-ER HAVE gun' with 'N-human-ER not HAVE gun' would be the criterion for eliminating terminal 2 as the appropriate element to which to link with. To arrive at these trees, some sophisticated form of language normalisation would be needed. This type of amphibology, then, is the point at which the tree's usefulness ceases and at which a grammar takes over.

The tree diagram suggested for the words "sucker" and "baste" is not meant to be the last word even in the resolution of ambiguity alone. As the representation of the vast body of encyclopedic knowledge to which all words have reference would be an immense undertaking, this thesis has necessarily been confined to formulating the questions that need to be asked rather than finding answers. For a computer programme the trees might be designated by one-dimensional bracket formulae of the type advocated in chapter 2. However, no attempt has been made to provide algorithms for them in this thesis, because the formulation of the information contained in the trees has not been based upon sufficient empirical evidence for the results of computation to decide its validity. The verification, therefore, of the procedure for constructing a computerised dictionary must await further study.

Footnotes to Chapter 4

1. Rangathan 1967
2. Tesnière 1959
3. Su 1971
4. Noel 1972
5. Fillmore 1969
6. Noel 1972
7. Wilks 1971
8. Schank 1969
9. Wilks 1971
10. Katz and Fodor 1963
11. Vickery 1965
12. Perry 1956
13. Dales 1963
14. Collier's Encyclopedia
15. Encyclopedia Americana
16. Harrap 1962
17. Wilks 1971
18. Booth, Brandwood and Cleave 1958
19. Masterman 1956
20. Bar-Hillel 1953
21. Micklesen 1961
22. Fillmore 1969
23. Smiles 1862
24. Gardin 1965
25. Katz and Fodor 1963

BIBLIOGRAPHY

Abraham, S. and Kiefer, F. A Theory of Structural Semantics. The Hague Paris: Mouton and Co., 1966.

Andrews, D. D., et al. "Recent Advances in Patent Office Searching: Steroid Compounds and ILAS". Information Systems in Documentation. Ed. J. H. Shera, A. Kent and J. W. Berry. New York: Interscience Press, 1957. 447-77.

Apresyan, J. "Analyse distributionelle des significations et champs sémantiques structurés". Langages. Ed. T. Todorov. Paris: Didier Larousse, mars 1966. 1, 44-74.

Apresyan, Yu, D., Melcuk and Zolkovsky "Semantics and Lexicography: Towards a New Type of Unilingual Dictionary". Studies in Syntax and Semantics. Ed. F. Kiefer. Dordrecht, Holland: D. Reidel Publishing Company. 10 (1969), 1-33.

Baille, A. and Roualt, J. Un Essai de formalisation de la sémantique des langues naturelles. G.2200 Centre National de la Recherche Scientifique, Centre d'Etudes pour la Traduction Automatique. Martin d'Hères-Isère, France, December 1966.

Bakewell, K. G. B. "The Universal Decimal Classification". Classification for Information Retrieval. Ed. K. G. B. Bakewell. London: C. Bingley, 1968.

Bar-Hillel, Y. "A Quasiarithmetical Notation for Syntactic Description". Language, 29 (1953), 47-58.

Barthes, R. "Essai de description d'un espace sémantique". Cahiers de lexicologie, 1 (1968), 15-36.

Bellert, I. "Arguments and Predicates in the Logico-Semantic Structure of Utterances". Studies in Syntax and Semantics. Ed. F. Kiefer. Dordrecht, Holland: D. Reidel Publishing Company, 10 (1969), 34-54.

Bierwisch, M. and Kiefer, F. "Remarks on Definitions in Natural Language". Studies in Syntax and Semantics. Ed. F. Kiefer. Dordrecht, Holland: D. Reidel Publishing Company, 10 (1969), 55-79.

Bobrow, D. G., Fraser, J. B. and Quillian, M. R. "Automated Language Processing". Annual Review of Information Science and Technology. Ed. C. A. Cuadra. New York: Interscience Pub., 1967. II, 161-86.

Bolinger, D. "The Atomisation of Meaning". Language, 41 (1965), 555-73.

Booth, A. D., Brandwood, L. and Cleave, J. P. Mechanical Resolution of Linguistic Problems, London: Butterworth's Scientific Publications, 1958.

Borko, H. "Indexing and Classification". Automated Language Processing. Ed. H. Borko. Santa Monica, Calif.: System Development Corporation, 1967, pp. 99-125.

Brisch, E. G. "Subject Analysis in Eighty-one Concepts". Aslib Proceedings, 7 (1955), 157-62.

Catford, J. C. A Linguistic Theory of Translation-Essay in Applied Linguistics. London: Oxford University Press, 1965.

Ceccato, S. Linguistic Analysis and Programming for Mechanical Translation. Technical Report No. RADC-TR-60-18. Milan, Italy: Giangiacomo Feltrinelli, 1960.

Ceccato, S. "Automatic Translation of Languages". Automatic Translation of Languages. Ed. S. Ceccato. Oxford: Pergamon Press, 1962, pp. 83-[illegible].

Chomsky, N. Aspects of the Theory of Syntax. Cambridge, Mass.: The M.I.T. Press, 1965.

Costello, J. C. Jr. "Uniterm Indexing Principles, Problems and Solutions". American Documentation, 12 (1961), 20-26.

Course in General Linguistics. Ed. Charles Bally and Albert Sechehaye. London: Peter Owen, 1959.

Coyaud, M. Introduction à l'étude des langages documentaires. Centre National de la Recherche Scientifique, Section d'Automatique Documentaire. University of Alabama Press, 1966.

Coyaud, M. et Siot-Decauville, N. L'Analyse Automatique des Documents. Paris et la Haye: Mouton and Co., 1967.

Curry, H. B. "Mathematics, Syntactics, and Logic". Mind, 62 (1953), 172-83.

Dales, R. Annelids. London: Hutchinson University Library, 1963.

Deese, J. The Structure of Association in Language and Thought. Baltimore: The John Hopkins Press, 1965.

Delavenay, E. An Introduction to Machine Translation. London: Thames and Hudson Ltd., 1960.

Dewey Decimal Classification and Relative Index. 17th ed. 2 vols. New York: Forest Press Inc., 1965.

Dostert, L. E. "The Georgetown-IBM Experiment". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth. Cambridge, Mass.: Technology Press of M.I.T. and John Wiley and Sons Inc., 1955, pp. 124-35.

Dubois, J. "Distribution, ensemble et marque dans le lexique". Cahiers de Lexicologie, 1 (1964), 5-16.

Duchaček, O. "L'antonymie" Cahiers de Lexicologie, 1 (1965), 55-66.

Earl, L. L., Bhimani, B. V. and Mitchell, R. P. "Statistics of Operationally Defined Homonyms of Elementary Words". Mechanical Translation, 10 (1967), 18-25.

Earl, L. L. and Robison, H. R. Automatic Information Abstracting and Extracting. Detroit: Management Information Services, 1970.

Edmundson, H. P. "Mathematical Models in Linguistics and Language Processing". Automated Language Processing. Ed. H. Borko. Santa Monica, Calif.: System Development Corporation, 1967, pp. 33-96.

Edmundson, H. P. and Wyllys, R. E. "Automatic Abstracting and Indexing-Survey and Recommendations". Communications of the Association for Computing Machinery, 4 (1961), 226-34.

Farradane, J. E. L. "A Scientific Theory of Classification and Indexing and its Practical Applications". Journal of Documentation, 6 (June 1950), 83-99.

Farradane, J. E. L. "A Scientific Theory of Classification and Indexing: Further Considerations". Journal of Documentation, 8 (June 1952), 73-92.

Fillmore, C. J. "The Case for Case" Universals in Linguistic Theory. Ed. E. Bach and R. T. Harms. New York: Holt Rinehart and Winston Inc., 1968, pp. 124-69.

Fillmore, C. J. "Types of Lexical Information". Studies in Syntax and Semantics, Dordrecht, Holland, 10 (1969), 107-37.

Gardin, J. C. "Four Codes for the Description of Artifacts; an Essay in Archaeological Technique and Theory". American Anthropologist, 60 (1958), 335-57.

Gardin, J. C. "SYNTOL". Systems for the Intellectual Organisation of Information II. New Brunswick, N.J.: Rutgers Series, 1965.

Garner, R. J. The Grafter's Handbook. London: Faber and Faber Ltd., 1947.

Goodenough, Ward H. "Componential Analysis and the Study of Meaning". Language, 32 (1956), 195-216.

Greenburg, J. H. Language Universals. The Hague: Mouton and Co., 1966.

Grolier, Eric de. A Study of General Categories Applicable to Classification and Coding in Documentation. Paris: Unesco, 1962.

Harper, K. E. "A Preliminary Study of Russian". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth. Cambridge, Mass.: Technology Press of M.I.T. and John Wiley and Sons Inc., 1955, pp. 66-85.

Harper, K. E. Proceedings of the National Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961, p. 423.

Harrap's Standard French and English Dictionary. Ed. J. E. Mansion. London: George C. Harrap and Company Ltd., 1962.

Harris, Z. S. String Analysis of Sentence Structure. The Hague: Mouton and Co., 1965.

Herdan, G. Language as Choice and Chance. Groningen: P. Noordhoff, 1956.

Hirschberg, L. "L'Utilisation de l'information sémantique dans le choix des unités lexicales dans les microglossaires". Rapport pour le Colloque de Nancy sur la Linguistique appliquée. Groupe de Linguistique Automatique, Université libre de Bruxelles, 1964 October 26-31.

Hubbs, Carl, L. and Lagler, Karl F. Fishes of the Great Lakes Region. Ann Arbor: The University of Michigan Press, 1958.

Jakobson, R. "On Linguistic Aspects of Translation". On Translation. Cambridge, Mass.: Harvard University Press, 1959, pp. 232-239.

Josselson, H. H. and Janiotis, A. "Multiple Meaning in Machine Translation". 1961 International Conference on Machine Translation of Languages and Applied Language Analysis. London: Her Majesty's Stationery Office, 1962. II, 406-15.

Kaplan, A. "An Experimental Study of Ambiguity and Context". Mechanical Translation, 2 (1955), 39-47.

Katz, J. J. and Fodor, J. A. "Structure of a Semantic Theory". Language, 39A (1963), 170-210.

Katz, J. J. and Fodor, J. A. "Structure d'une théorie sémantique avec applications au Francais". Cahiers de Lexicologie, 2 (1966), 39-72.

King, G. Proceedings of the National Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961, pp. 53-62.

Kuno, S. "The Predictive Analyzer and a Path Elimination Technique". Readings in Automatic Language Processing. Ed. D. G. Hays. New York: American Elsevier Publishing Company Inc., 1966, pp. 83-106.

Kuroda, S. -Y. "Remarks on Selectional Restrictions and Presuppositions". Studies in Syntax and Semantics, Dordrecht, Holland, 10 (1969), 138-67.

Laffal, Julius. "Towards a Conceptual Grammar and Lexicon". Computers and the Humanities, 4 (January 1970), 173-87.

Lambek, J. "The Mathematics of Sentence Structure". American Mathematical Monthly, 65 (1958), 154-70.

Ledley, R. S. "Tabledex: A New Coordinate Indexing Method for Bound Book Form Bibliographies". Proceedings of the International Conference on Scientific Information, 2 (1959), 1221-43.

Leech, G. N. Towards a Semantic Description of English. London: Longman's, Green and Co. Ltd., 1969.

Lees, R. B. "Structural Grammars". Mechanical Translation, 4 (1957), 5-10.

Lukjanow, A. W. "Report on Some Principles of the Unified Transfer System". Proceedings of the National Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961a, pp. 88-120.

Lukjanow, A. W. "Semantic Classification". Proceedings of the National Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961b, pp. 394-97.

Lyons, John. Structural Semantics, an Analysis of Part of the Vocabulary of Plato. Oxford: Basil Blackwell, 1963.

Mann, K. H. Leeches (Hirudinea) Their Structure, Physiology, Ecology and Embryology. Oxford, London: Pergamon Press, 1962.

Martinet, A. Elements of General Linguistics. London: Faber and Faber Ltd., 1964.

Masterman, M. "New Techniques for Analysing Sentence Patterns". Mechanical Translation, 3 (1956), 4-6.

Masterman, M. "The Thesaurus in Syntax and Semantics". Mechanical Translation, 4 (1957), 35-44.

Masterman, M., Needham, R. M. and Sparck-Jones. "The Analogy between Mechanical Translation and Information Retrieval". Proceedings of the International Conference on Scientific Information. 2 (1959), 917-35.

Masterman, M., Parker-Rhodes, A. F., Richens, R. H. and Halliday, M. A. K. "Report on Research at the C.L.R.U.". Mechanical Translation, 3 (1958), 36-7.

Matthews, G. H. and Rogovin, S. "German Sentence Recognition". Mechanical Translation, 5 (1958), 114-20.

McCawley, J. D. "The Role of Semantics in a Grammar". Universals in Linguistic Theory. Ed. E. Bach and R. T. Harms. New York: Holt Rinehart and Winston Inc., 1968, pp. 125-69.

Melton, Jessica. "Procedures for Preparation of Abstracts for Encoding". Tools for Machine Literature Searching. Ed. A. Kent, J. W. Perry and J. L. Melton. New York, London: Interscience Publishers, 1 (1958), 69-150.

Melton, J. L. "The Semantic Code". Tools for Machine Literature Searching. Ed. A. Kent, J. W. Perry and J. L. Melton. New York, London: Interscience Publishers, 1 (1958), 221-79.

Melton, J. and Perry, J. W. "Introduction to Analysis of Questions". Tools for Machine Literature Searching. Ed. A. Kent, J. W. Perry and J. L. Melton. New York, London: Interscience Publishers, 1 (1958), 381-456.

Mersel, J. "Research in Machine Translation at Ramo-Wooldridge". Proceedings of the National Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961, pp. 26-38.

Micklesen, L. R. "An Experiment in the Automatic Selection or Rejection of Technical Terms". Proceedings of the National Symposium on Machine Translation. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961, pp. 398-403.

Mills, Jack. "The Universal Decimal Classification". Systems for the Intellectual Organisation of Information I. New Brunswick, N.J.: Rutgers Series, 1964.

Mooers, Calvin, N. "Zatocoding and Developments in Information Retrieval". Aslib Proceedings, 8 (1955), 3-22.

Mosteller, F. and Wallace, D. L. Inference and Disputed Authorship: the Federalist. Reading, Mass.: Addison Wesley Company Inc., 1964.

Mounin, G. La Machine à traduire. London: Mouton and Co., 1964.

Mounin, G. "Essai sur la structuration du lexique de l'habitation". Cahiers de Lexicologie. 1 (1965), 9-24.

Müller, C. Initiation à la Statistique Linguistique. Paris: Hollier-Larousse, 1968

Nichols, David. Echinoderms. London: Hutchinson University Library, 1962.

Nida, E. A. "Principles of Translation as Exemplified by Bible Translating". On Translation. Ed. R. Jakobson. Cambridge, Mass.: Harvard University Press, 1959, pp. 11-31.

Nida, E. A. Toward a Science of Translating. Leiden, Netherlands: E. J. Brill, 1963.

Noël, J. Linguistic Problems in Mechanised Indexing of English Abstracts. University of Victoria, May 1968.

Noël, J. A Semantic Analysis of Abstracts. Around an Experiment in Mechanised Indexing, part III, 1972.

Oettinger, A. G. Automatic Language Translation. Cambridge, Mass.: Harvard University Press, 1960.

Oxford English Dictionary (unabridged). Ed. A. H. Murray, H. Bradley, W. A. Craigie, C. T. Onions. Oxford: Clarendon Press, 1933.

Parker-Rhodes, A. F. "The Use of Statistics in Language Research", Mechanical Translation, 5 (1958), 67-73.

Parker-Rhodes, A. F. "Some Recent Work on Thesauric and Interlingual Methods in Machine Translation". Information Retrieval and Machine Translation. Ed. A. Kent. New York, London: Interscience Publishers Inc., 1961, pp. 923-934.

Perry, J. W. "Translation of Russian Technical Literature by Machine". Mechanical Translation, 1 (1955), 15-24.

Perry, J. W., Kent, A. and Berry, M. M. Machine Literature Searching. New York, London: Western Reserve University Press, Interscience Publishers, 1956.

Pimsleur, P. "Semantic Frequency Counts". Mechanical Translation, 4 (1957), 11-13.

Postal, P. M. "Underlying and Superficial Linguistic Structure". Harvard Educational Review, 34 (1964), 246-66.

Pottier, B. Systématique des Eléments de Relation. Etude de Morphosyntaxe Structurale Romane. Paris: Klincksieck, Série A: Manuels et Etudes linguistiques-2-, 1962.

Prieto, L. J. Principes de noologie. Fondements de la théorie fonctionelle du signifié. Londres--La Haye--Paris: Mouton and Co., 1964.

Quillian, M. R. "Computers in Behavioural Science. Word Concepts: A Theory and Simulation of some Basic Semantic Capabilities". Behavioural Science. Ed. J. G. Miller. 12 (1967), 410-30.

Random House Dictionary of the English Language (Unabridged). Ed. Jess Stein. New York: Random House, 1966.

Rangathan, S. R. Prolegomena to Library Classification. London: Asia Publishing House, 1967.

Rees, Thomas H. Jr. "Standardised Telegraphic Abstracts from Articles in New York Times". Tools for Machine Literature Searching. Ed. A. Kent, J. W. Perry and J. L. Melton. New York, London: Interscience Publishers. 1 (1958), 69-150.

Reifler, E. "Report on the First Conference on M.T." Mechanical Translation, 1 (1954), 23-32.

Reifler, E. "The Mechanical Determination of Meaning". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth. Cambridge, Mass.: Technology Press of M.I.T. and John Wiley and Sons Inc., 1955, pp. 136-164.

Reifler, E. "The Solution of MT Linguistic Problems through Lexicography". Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961a, pp. 312-16.

Reifler, E. "Current Research at the University of Washington". Symposium on Machine Translation. Ed. H. P. Edmundson. Englewood Cliffs, N.J.: Prentice Hall Inc., 1961b, pp. 155-59.

Richens, R. H. "Preprogramming for M.T.". Mechanical Translation, 3 (1956), 20-28.

Richens, R. H. and Booth, A. D. "Some Methods of Mechanical Translation". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth. Cambridge, Mass.: Technology Press of M.I.T. and John Wiley and Sons Inc., 1955, pp. 124-35.

Salton, G. The Identification of Document Content: A Problem in Automatic Information Retrieval. Brookline: Harvard Symposium on Digital Computers and their Applications, American Academy of Arts and Sciences, April 1961.

Salton, G. "Manipulation of Trees in Information Retrieval". Communications of the ACM, 5 (1962), 103-14.

Salton, G. "Automatic Phrase Matching". Readings in Automatic Language Processing, Ed. D. G. Hays. New York: American Elsevier Publishing Company Inc., 1966, pp. 169-88.

Sayers, W. C. Berwick. A Manual of Classification for Librarians. London: Andre Deutsch Ltd., 1967.

Schank, R. C. A Conceptual Dependency Representation for a Computer-Orientated Semantics, University of Texas, 1969.

Schultz, Claire K. H. P. Luhn, Pioneer of Information Science, Selected Works. New York: Spartan Books, London: Macmillan and Co. Ltd., 1968.

Schmidt, Gerald D. How to Know the Tapeworms. Dubuque, Iowa: W. M. C. Brown Company Publishers, 1970.

Sharp, J. R. Some Fundamentals of Information Retrieval. London: Andre Deutsch, 1965.

Simmons, R. F. "Automated Language Processing". Annual Review of Information Science and Technology, 1 (1966), 137-69.

Slama-Cazacu, T. Langage et Contexte. Le Problème du Langage dans la Conception de l'Expression et de l'Interpretation par des Organisations Contextuelles, 's-Gravenhage: Mouton and Co., 1961.

Smiles, Samuel. Lives of the Engineers, London: John Murray. 3 (1862), page 10.

Solomonoff, R. J. "A Progress Report on Machines to Translate Languages and Retrieve Information". Advances in Documentation and Library Science. Ed. A. Kent. New York: Interscience Publishers. 3 (1961), 941-53.

State of the Library Art. Ed. R. R. Shaw. New Brunswick, N.J.: Graduate School of Library Science-The State University, 1961.

Stevens, M. E. Automatic Indexing: A State-of-the-Art Report, Washington, D.C.: National Bureau of Standards Monograph 91, 1965, March 30.

Stindlova, J. "Les Dictionnaires Inverses". Cahiers de Lexicologie, 2 (1960), 79-86.

Su, Y. W. A Computational Model of Paragraph Production. University of Florida, Gainesville: Technical Report No. 71-102, Centre for Informatics Research, 1971 November.

Taube, Mortimer et al. Studies in Coordinate Indexing, 3 vols. Washington, D.C.: Documentation Incorporated, 1953-56.

Taube, Mortimer et al. "Notes on the Use of Roles and Links in Coordinate Indexing". American Documentation, 12 (1961), 98-100.

Tesnière, Lucien. Eléments de Syntaxe Structurale. Paris: librairie C. Klincksieck, 1959.

Todorov, T. "Recherches Sémantiques". Langages. Ed. T. Todorov. Paris: Didier Larousse, mars 1966, pp. 5-43.

Uldall, H. J. "Notes on the English Tenses". English Language Teaching, 2 (1948), 122-28 and 147-53.

Ulvestad, B. "Syntactical Variants". Mechanical Translation, 4 (1957), 28-34.

Use of the Library of Congress Classification. Ed. R. H. Schimmelpfeng and C. D. Cook. New York: American Library Association, 1966 July 7-9.

Van Buren, P. "Mechanisation in Lexis". Cahiers de Lexicologie, 1 (1968), 71-84.

Vickery, B. C. Classification and Indexing in Science (2nd ed.). London: Butterworth's, 1959.

Vickery, B. C. On Retrieval System Theory (2nd ed.). London: Archon Books, Butterworth's, 1965.

Vinay, Jean-Paul. "Codage automatique et dictionnaires". Mémoires de la Societé Royale du Canada, série IV, Tome VII, 1969.

Wadington, John P. "Unit Term Coordinate Indexing". American Documentation, 2 (1958), 107-13.

Wallace, E. C. "Use of the ASM-SLA Metallurgical Literature Classification System by Industrial Metallurgists". Special Libraries, 47 (March 1956), 114-16.

Weaver, W. "Translation". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth. Cambridge, Mass.: Technology Press of M.I.T. and John Wiley and Sons Inc., 1955, pp. 15-23.

Webster's Third New International Dictionary (Unabridged). Ed. P. B. Grove. Springfield, Mass.: G. and C. Merriam Company, 1965.

Weinreich, U. "Travels through Semantic Space". Word, 14 (1958), 346-66.

Weinreich, U. "Explorations in Semantic Theory". Current Trends in Linguistics, 3 (395-477), 1966.

Wilks, Yorick. "An Artificial Intelligence Approach to Machine Translation". Artificial Intelligence Project, Draft 8, Stanford University, Calif., 1971 December.

World Book Encyclopedia Dictionary. Ed. C. L. Barnhart. Chicago: Doubleday and Company, 1965.

Wyllys, R. E. "Extracting and Abstracting by Computer". Automated Language Processing. Ed. H. Borko, New York: Wiley, 1967, pp. 127-79.

Yngve, V. H. "Syntax and the Problem of Multimeaning". Machine Translation of Languages. Ed. W. N. Locke and A. D. Booth, 1955a, pp. 208-26.

Yngve, V. H. "Sentence-for-Sentence Translation". Mechanical Translation, 2 (1955b), 29-37.

Yngve, V. H. "Mechanical Translation Research at MIT". Mechanical Translation, 3 (1956), 44-45.

Yngve, V. H. "A Framework for Syntactic Translation". Mechanical Translation, 4 (1957), 59-65.

Yngve, V. H. "The Feasability of Machine Searching of English Texts". Proceedings of the International Conference on Scientific Information, 2 (1958a), 975-95.

Yngve, V. H. "A Programming Language for Mechanical Translation". Mechanical Translation, 5 (1958b), 25-41.